THREE FACTOR DIFFERENTIAL ITEM FUNCTIONING ANALYSIS OF A PROVINCIAL FINAL EXAMINATION

by

Willy Chow

B.Sc., University of British Columbia, 1983

PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF EDUCATION

in

CURRICULUM AND INSTRUCTION

THE UNIVERSITY OF NORTHERN BRITISH COLUMBIA

August 2003

Abstract

Assessment methods should be developed or chosen so that the inferences drawn about the knowledge, skill, attitudes, and behaviors possessed by each student are valid and not open to interpretation. Differential item functioning analysis can be used to monitor the validity and fairness of examinations. A differential item functioning analysis was performed on the January 1999 and June 1999 sittings of the Chemistry 12 provincial final examination. Gender, school district size and minority group membership were the three parameters of this investigation. Each of the items that comprised the examination were subject to Rasch analysis to generate difficulty estimates for our reference and comparison groups. The open-ended items were further subject to a partial credit analysis. Various levels of DIF were found: items were found to show DIF, and subtests were found to show DIF or no net DIF. If DIF was present, it was limited to no more than two items. The items discriminated equitably between the reference and comparison groups for the three factors of study. Both sittings of the examination were found to be valid for our comparison groupings.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One

Introduction to Problem

For a test to be useful and acceptable, it must be fair to all sub-groups of the population for which it is to be the instrument of measurement. Generally the role of a test is to discriminate among the participants on the traits or abilities the test was designed to measure. The test should not discriminate between sub-groups on any other basis that is unrelated to the purpose of the test. There should be no unfair advantage to any sub-group based on attributes such as gender, social class or ethnic group. A test is unfair, or more precisely, it produces invalid results, if it is not as accurate for one sub-group as it is for the rest of the test population. The validity of the test is different for the two groups.

Differential item functioning (DIF) occurs when different groups of examinees show differing probabilities of success on (or endorsing) items after matching on the underlying ability that the items are intended to measure (Zumbo, 1999). The DIF analysis is a procedure used to determine if test questions are fair and appropriate for assessing the knowledge of various comparison groups that make up the population. It is based on the assumption that test takers who have similar knowledge that is based on test scores should perform in similar ways on individual test items regardless of their demographic membership. When individuals with similar abilities do not perform in a similar fashion on a test item, that item is said to display DIF. The presence of DIF in a test is a serious problem affecting the validity of the item as well as the validity of the entire test. Patterns of differential item functioning that suggest actual group differences have staggering implications for policy makers, educators and curriculum developers.

The DIF analysis for items is important in test development because it helps examine and eliminate items that may be potentially unfair to sub-populations due to cultural or gender differences or to some other group membership. Identifying the factors that are associated with DIF would significantly contribute to the development of valid assessment instruments. Differential item functioning analyses do not ensure that item content is valid. The construct validity of each item still needs to be confirmed. The DIF analyses are also separate from other validation studies (Clauser & Mazor, 1998). Procedures for DIF detection are designed to identify individual items that function differentially relative to some identified criterion. The meaningful interpretation of DIF statistics presupposes appropriate construct and predictive validity evidence.

Checking for DIF has become a routine practice for both large and small standardized testing programs. This exercise exposes items that favor one subgroup over others due to characteristics that typically are extraneous to the attributes being tested. According to Pashley (1992), impact studies are insufficient unless average group performances are known to be equal. Since subgroups are not always well represented across the entire ability scale, matching samples, as required by most observed score analyses, can be a problem. Item response theory provides a method that controls for these ability differences and eliminates this problem.

All those who construct tests should be sensitive both to the critical role that the testing plays and to the different types of diversity of the test takers. Test makers should be committed to reviewing their tests and to ensure that all their tests are fair for examinees regardless of group membership. On a larger scale, assessment systems should also be regularly reviewed and improved to ensure that they are educationally beneficial for all students.

Wright, Mead, and Draba (1976) state that successful development of the proposed procedures for detecting and correcting bias will have implications for both pupil evaluations and measurement research. These procedures will allow practitioners to detect biased items, identify what defines the intended trait for all groups and evaluate the test protocol of every person with respect to bias.

Possible Factors

According to Clauser and Mazor (1998), differential item functioning is present when examinees from different groups have different probabilities or likelihood of success on an item after they have been matched on the ability of interest. For the purpose of this study, DIF will be consider for groups of persons that are determined by gender, school district size, and minority group membership. More specifically, attempts to measure gender differences have never occurred so frequently. Differential item functioning offers a method of determining if the individual items that collectively define these measures are valid for both males and females. School district size is a complex variable that is known to be associated with many differences that are relevant for those who attempt to study education. Some of these differences include the socio-economic status of students, the budget of schools, the urban versus rural nature of schooling, and, ultimately, the educational experiences of students. Do students from larger school districts outperform students from smaller school districts? The answer to that question assumes that the items of the measure used to determine performance are not differentially discriminating for or against the many groups of students that are known to be associated with school districts of different size. The third variable, minority group membership, will focus upon comparisons of the academic performance of aboriginal and non-aboriginal students. When differences between aboriginals and non-aboriginals are reported, do the items in the measures that have detected those differences really measure differences in performance?

A great many explanations have been offered for the differences in test performance among various population subgroups. In fact, so many explanations have been offered that to date, the results are still largely inconclusive (Scheuneman & Slaughter, 1991). This failure leaves one without the knowledge necessary to put test performance in a proper perspective. This point can be illustrated by considering in greater detail the challenge of determining valid differences in gender, school district size, and race.

Gender

Gender is a demographic variable that has been scrutinized long before DIF analyses came into being. The problem of gender bias has been well documented for many years. There may be differences and similarities in the innate intellectual function of females and males but opinions vary widely on the matter. For example, Borich & Tombari (1995, p. 613) concluded. "males and females are born with similar spatial, mathematical and verbal ability." Several studies (Dillon, 1982; Haggerty, 1991; Ma, 1995) suggest that gender differences are negligible in specific realms. It is commonly viewed as a misconception that one's gender is a significant predictor of abilities and interests (Campbell & Storo, 1996). From this research, there should not be any differences attributed to gender based on innate intellectual functioning.

Refuting studies (Gambell & Hunter, 1997; Hativa, 1989; Schofield, 1982; Tocci & Engehard, 1991) suggest significant differences in the same realms. Cognitive sex differences are thought to reflect differences in information processing. "Males are considered to be more likely to organize information in a self related manner whereas females are more likely to adopt a comprehensive approach to information processing: (McGivern, Huston, et al, 1997, p. 323). The differences in performance may be attributed to gender differences in innate intellectual functioning.

Regardless of the research in innate intellectual functioning, socialization plays a part in the differences that are observed in the classroom. "Performance differences between the sexes are for the most part learned behaviors induced by societal expectations and the behavior of adults" (Good & Brophy, 1991, p. 29). Schools may have perpetuated under-achievement. Social and religious traditions also contribute to these inequities.

Public accountability pressures and the resulting need to demonstrate educational quality have provided educators and researchers with access to evidence about gender difference and its relationship to educational system outcomes. This evidence indicates that the relationship is far from a simple one. According to the Women's Freedom Network (1998), females do better than males in some areas and males do better than females in some other areas. To better understand the relationship as a whole it is necessary to examine the advantages and disadvantages of both genders:

> Females lag behind males in two academic areas: mathematics and science achievement. Females also lag slightly behind males in attaining professional, business and doctoral degrees. Conversely, males lag behind females in two other academic areas and by far wider margins. These areas are reading achievement and writing skills. Males are far more apt than females to end up at the bottom of the class in school and to be placed in special classes for students with learning disabilities. Males believe that the school climate is hostile toward them in comparison to females. Relatively speaking, females do not. Males receive less encouragement than females and males perceive that less is expected from them in comparison to females. (p. 1)

For the past twenty years, the North American educational system has been preoccupied with equalizing opportunity for female students because of inequity in female prospects in the job and post-secondary education markets (Gambell & Hunter, 1997). According to Edge, Fisher, Martin and Morris (1997), strategies were implemented to promote gender equity within the classroom, heighten awareness of female contributions to society, increase teacher understanding of the consequences of gender inequity and heighten awareness among students of the existing problem. To identify the inequities, researchers have examined local communities, the home environments, textbooks, modes of instruction in the classroom and student attitudes, among other factors.

According to Hoff Sommers2000), females outshine males at school. Females get better grades and have higher educational aspirations. Females follow more rigorous academic programs and participate in advanced placement classes at higher rates. Hoff notes from the National Center for Educational Statistics, that slightly more girls than boys enroll in high level math and science courses and that females are more academically engaged than their male counterparts. Hoff Sommers concluded from examining data from numerous Western countries that the disadvantaged gender in our schools is the male gender. Boys - not girls - now suffer from more learning disabilities and attain fewer post-secondary degrees (Hoff Sommers, 2000).

Casting aside physical differences, males and females have been compared in just about every imaginable arena. Their academic achievement has been monitored over the year in response to the accountability of educational reform. Female resurgence coupled to a male decline has been observed for their groups. Some researchers contend that females and males are born with similar innate functioning, the differences must be the result of different interactions with their environments. Even though schools may not be the sole cause of gender differences in achievement, the schools still have an important role to play in making sure that both girls and boys have appropriate and equal opportunities to develop their intellectual skills.

The research literature on gender differences in scores on cognitive tests and the origins of these differences are both complex and contentious. In the general population, most gender differences on standardized tests of achievement are small and negligible (Women's Freedom Network, 1998). This may be due to the fact that group characteristics are difficult to eliminate from general population comparisons. Differential item functioning, addresses this problem. A DIF analysis permits investigators to identify the differences that are specific to individual items in the measures and separate them from the residual differences in performance that are the focus of these measures.

School District Size

The second variable, school district size, is also complex. According to McGuire (1989), the school district may not be the appropriate unit of analysis for researchers. Examination at the school district level may amplify the fact that size economies bring many constraints. School district size is an even more elusive variable to judge since the nature and the mission of the school district is not uniformly defined. Walberg and Fowler (1987) found that high socioeconomic status school districts achieved more than did lower socioeconomic status school districts. Howley (1994) found when all else is held equal (particularly community or individual socioeconomic status); comparisons of schools or districts based on differences in enrollment generally favor smaller units.

The size of schools is positively correlated with the population of school districts. Large school districts tend to have large schools, while small school districts tend to have small schools, many of which are classified as rural. According to Fowler (1992), sufficient theoretical work has been done to suggest how the mechanism of school size affects student outcomes. Secondary school size determines student satisfaction with academic courses, attendance, and participation in extracurricular activities. Student achievement is enhanced by satisfaction with academic courses, a low dropout rate, and voluntary participation in extracurricular areas. All of the characteristics are frequently found to be characteristic of small secondary schools.

The essence of the problem is the mean school size (Coleman & Laroque, 1986). Small schools in rural areas cannot possibly offer the same educational experiences that can and are offered by large school, which are typically located in urban centers. In addition to school size affecting student affective and cognitive outcomes, high school size is related to curricular offerings (Fowler, 1992). Generally speaking, the number and variety of course offerings are positively correlated with the enrollment of high schools. Large schools have more students with similar needs, and thus are better able to create specialized programs to address those needs: thus, they are better able to create specialized programs to address those needs than are small schools. In contrast, small schools must focus resources on core programs by excluding marginal students from programs or absorbing them into programs that may not meet their needs (Lee & Smith, 1996). Roellke (1996) states that small high schools face the challenge of maintaining a broad curriculum while seeking to offer more advanced courses such as calculus or even Advanced Placement programs.

Unequal access to a variety of facilities and experiences that enhance an individual's knowledge base may lessen the performance of students in smaller schools. Differences in educational experiences as explanations of these deficits includes three types of discussion: the number and quality of courses taken by specific groups, the quality of teachers and the teaching given to the various groups, and the motivation of the students as it relates to their experiences within the educational environment. This discussion is based on environmental factors. Environmental factors including district size, fiscal resources, percentage of non-white students in the population of the district's community, and the education and income levels of the parental populations. All of these factors influence the effectiveness of school and the achievement of students (Bidwell & Kasarda, 1975).

Howley and Harmon (1999) argue that small schools usually flourish. They are more productive and effective than are larger schools. Furthermore, their students make more rapid progress toward graduation, are more satisfied, drop out less frequently, and behave better than do students in large schools. Yet in many rural parts of the country, public officials and professional educators continue to believe that small schools are inefficient and ineffective. Rural communities have seen this way of thinking result in closed schools and long bus rides for many students. Educators who are entrusted to make the best decisions on behalf of students have to balance the economics of the situation with student success.

Monk (1992) notes that existing research on school and school district size is not as conclusive as policy makers might wish. A large school or school district does not guarantee desirable results. Recommended school sizes have been declining with recent efforts to restructure education by emphasizing school autonomy, local decision making and the development of schools within schools. Walberg (1992) cites literature that supports the argument that small districts produce better student test scores than do large districts. High achieving regions frequently have small districts and small schools.

One of the reasons that the variable of school district size is so complex, that it is dependent on population and conditions that are related to population density. The population base that the student population comes from will influence the choices students make, the education values they hold and the educational values that motivate them. Student population also determines school organization and the availability of curricular offerings within the district. Students from smaller locales are disadvantaged compared to their peers from larger districts.

Even though the size of school districts may not be the lone cause of differences in student achievement, it is important because the school districts provide opportunities for students to develop their intellectual skills. The educational experience in small and large districts may not be equal. Again, differential item functioning offers the possibility of becoming a method of analysis that will help answer this question. It offers the possibility of ascertaining the examinee characteristics associated with students from small and large schools or small and large school districts.

Minority Group Membership: Ethnicity

Another factor that has been considered in differential item functioning is ethnicity or minority group membership. The majority group versus minority group comparison is common on formal assessments such as the Scholastic Aptitude Test. Allen and Wainer (1989) contend that the accuracy of the procedures used to compare the performance of different groups of examinees on test items depends on the correct classification of each examinee group. The significance of this dependence is determined by the sensitivity of the statistical procedure utilized and the proportion of the examinees that are not classified correctly. Their study found that efforts to obtain more accurate ethnic identification of the examinees were rewarded by using DIF analyses to improve the accuracy of the classification. The examinees for whom ethnicity was not specified were found to contribute to significant changes in DIF measures.

The specific ethnic group to be considered in this study is aboriginal persons. Aboriginals are defined as Indians, status and non-status, Metis and Inuit (Malatest, Barry, Krebs & Whyte, 2002). The literature is sparse in this area and it is politically sensitive. Only recently have data been collected on the participation of aboriginal students on formal assessments. Data such as dropout rates lead one to conclude that provincial examinations are not priorities with this group. These students have other extenuating factors that limit their participation; for example, many come from communities that are challenged by poverty. A Canadian study completed by the Department of Indian and Northern Affairs (1997) reported that the educational outcomes of aboriginal and non-aboriginal students from socio-economically equivalent communities were all but equal in most respects except one. The dropout rate in aboriginal communities was very much greater than it was in non-aboriginal communities. Some aboriginal students have participated in some formal assessments but their results are often masked or are blended into the group statistics. Researchers have not been privy to these results. Consequently, only policy makers and appropriate government agencies are aware of the results.

Because research on the academic performance of aboriginals in British Columbia is limited, it is necessary to review the research on the performance of other aboriginal groups within North America to obtain information that may be relevant for British Columbia students. Two studies completed by American investigators suggest that the academic success and participation rates of Native Americans are similar to those of Native Canadians. Riles (1995) reported a school dropout rate of 29.2% for a sample of Native American students. A similar result for Native American students is also reported in a cross-cultural study by Hanson and Farrell (1995). In this study, it was noted that all the students who learned to read in kindergarten were subsequently found to be superior in reading skills and in all other education indicators as seniors in high school. Regrettably, the gains made by Native American students were significantly lower than those of the students from all the other minority groups. Specifically, the values of the gains reported are: Blacks - 13.5%, Asians - 10.5%, Hispanics - 7.0% and Native Americans - 3.0%.

Instead of blaming social and economic factors that disadvantage many aboriginal students, Reyhner (1992) turns the dropout focus to the failings of the school system. In particular, he contends that the blame should be attributed to large schools, uncaring and untrained teachers, passive teaching methods, inappropriate curricula, inappropriate assessment procedures, tracked classes and lack of parental involvement. The academic success of Native Americans needs to be nurtured. In Native American communities (American Indians, Alaska Natives, and Native Hawaiians), there must be lifelong learning opportunities that allow all Native Americans the opportunity to meet their tribal responsibilities and improve their quality of life.

Both of the studies noted above suggest that the academic difficulties of aboriginal students are not unique to Canada or British Columbia. Aboriginal students are struggling in the education system of both countries. At all grade levels and on all academic measures, aboriginal students are receiving failing grades. Surprisingly, a smaller proportion of aboriginal than non-aboriginal students are enrolled in special education programs or learning assistance programs. Conversely, a greater percentage of aboriginal students are enrolled in non-academic programs (Marx & Grieve, 1988). In Canada, a study of 36 British Columbia secondary schools found that schools with high aboriginal enrollment had higher dropout rates, lower graduation rates, and lower participation rates on Grade 12 government examinations than did schools with lower aboriginal enrollment (Cameron, 1990).

Aboriginal students are not successful in school due to a complex set of factors. Many of those factors are similar to those any other unsuccessful student would face. The continuing inability of aboriginal students to succeed within the curriculum context of a public education system is related to the effects of cultural dissonance, racial stereotyping, economic poverty, lack of school success models within families and the perceived lack of career opportunities.

Improving school success for aboriginal students has been a focus for local policy makers in the province of British Columbia. More (1998) has argued that aboriginal learners learn better if learning styles are compatible with aboriginal family practices, cultural traditions and ways of life. Changes that modify the ways schools and teachers respond to academic and social problems would be a start. Programs aimed at improving cross-cultural understanding between schools and aboriginal students and their families should receive consideration.

As a group, aboriginal students have high drop out rates and low graduation rates because most aboriginals choose not to participate in academic pursuits. Due to the low participation rates, the test scores lack any ability to distinguish properly differences in the ability levels of the aboriginal and non-aboriginal groups. Measures implemented to improve school success are difficult to substantiate. Differential item functioning analysis can accommodate the small subpopulations. A DIF analysis controls for the group differences which, in turn, permits the examination of residual differences in the performance of items that comprise the test. Differences in performance between aboriginals and non-aboriginals can then be monitored overtime.

Differential item functioning analysis serves to confirm that a test is fair to all applicants regardless of their ethnic group membership. Any assessment should be fair to all students. Bias found in the assessment instrument would indicate an unfair advantage for some students. The presence of DIF in a test is a serious problem affecting the validity of the item as well as the entire test. Poor assessment could possibly lead to limitations being placed on the future educational opportunities of individual students and groups of minority students.

The Model

Methods based on item response theory (IRT) provide a useful theoretical framework for DIF because between-group differences in the item parameters for the specific model can be used to model DIF. All the various DIF methods conceptualize DIF in terms of differences in the model parameters for the comparison groups (Clauser & Mazor, 1998). The general framework involves estimating item parameters separately for the groups. For detailed mathematical descriptions of item-response modeling, interested parties may refer to Hambleton and Swaminathan (1985), Lord (1980), or Weiss and Yoes (1991). The general framework involves estimating item parameters separately for the reference and focal groups. After placing the item parameters on the same scale, differences between the item parameters for the two groups can be compared. When the parameters are equal for the two groups, the item does not display DIF. When DIF is absent, the two groups' item characteristic curves (ICCs) overlap. When DIF is present, items may differ across groups in their difficulty or items may differ across groups in their discrimination and/or pseudo-guessing.

Utilizing the one-parameter model which is more commonly known as the Rasch model (Rasch, 1960; Wright & Stone, 1979), DIF is manifested as a difference in the difficulty parameter for each item, as the Rasch model works only with the difficulty parameter. The Rasch approach models the difference between examinee ability and item difficulty. In the Rasch model, the discrimination and pseudo-guessing parameters are sample dependent artifacts. Limiting the number of parameters used to characterize an item characteristic curve has the effect of making the mode more stable and elegant (Pope, 1998). Data that do not fit the Rasch model are considered poor data and are discarded from the analysis. It is the assumption of the Rasch model that all the data must fit the model. If they do not, those data are considered troublesome and are discarded.

Using the other two models, the two-parameter or three-parameter logistic model, DIF is quantified by the area between the item characteristic curves. Comparing item parameters may result in significant differences in item parameters but item characteristic curves not differing by more than 0.05 in the specified ability range (Linn, Levine, Hastings, and Wardrop, 1981). Raju (1990) has derived expressions for determining the area between the curves. The expressions function well for the two-parameter model but not for the three-parameter model if the pseudo-guessing parameter is not equal for the two groups. The gain of fitting more data is offset by the loss of the ability to calculate the significance test for the area. Consequently, the Rasch model will be the framework that will be utilized to check for DIF.

Chapter Two

Method

Background to the Examination

The current British Columbia Provincial Examination Program was established in 1984. It was implemented to ensure that students enrolled in academic subjects met consistent provincial standards of achievement; it also served to respond to the strong public desire for improved standards in education. Provincial examinations are developed by teachers who are contracted by the Ministry of Education. The examinations are based on the provincial curricula. The formal assessment provides useful information about whether students as a group have reached important learning goals. Provincial examinations are part of the graduation requirement in British Columbia. Provincial examinations are compulsory in English (literature and composition) for every graduating student. Other provincial examinations are written depending on the course selection of the individual student. Examinations are written for the traditional academic courses; possible examinations include Biology, Chemistry, Mathematics, French, and Geography.

These examinations are an integral component for the evaluation of each student. The examination makes up forty percent of the student's overall grade while the other sixty percent is decided at the school level. Since the examination carries such a large weight, the result will treat students fairly when applying for admission to universities and other post-secondary institutions as well as scholarships.

Participants

Those students who wrote the Chemistry 12 provincial examination during the January 1999 and June 1999 administration provided the data to be for analysis. The majority of the respondents were residents of British Columbia but a few were from the Yukon Territory. The pool of respondents included students from the public and private sectors. The January and June sittings were utilized because of the large number of students who write the provincial examination at those times. Other sittings occur in April and August, but those sittings were not utilized since the data pool was significantly smaller.

Demographic information about the respondents allowed for three groupings to be considered in this study. Two variables, gender and aboriginal status, are demographic data that the respondents volunteer upon registration for the examination. The test population has been divided into male and female respondents for the consideration of gender while the test population was divided into non-aboriginals and aboriginals for the second variable. Aboriginality is a self-reported variable. Some of the respondents may have not indicated that they were of aboriginal descent and been improperly included as non-aboriginals. Also included in the registration for each student was the school district in which the student was enrolled. School district membership was used to formulate a third variable of school district size. School districts with at least one community with a population of 100 000 or more were classified as large school districts. Those with a population less than 25 000 were categorized as small districts and those with a population between these values were classified as medium sized districts (see Appendix A).

Instrumentation

The basic format of the provincial examination has remained unchanged for years. The format is common across the examinations that are offered to those students enrolled in provincially examinable courses. The Chemistry examination consists of two components: a multiple choice component consisting of forty-eight items worth sixty percent and an open-ended written component varying in the number of items worth forty percent. Each examination is created using a table of specifications that outlines the curriculum organizers, suborganizers and cognitive level emphases (see Appendix B).

The multiple choice component and open-ended component reflect in their composition the curricular emphases for Chemistry 12. Reaction Kinetics, Dynamic Equilibrium, and Solubility Equilibria each make up 12.5% while Acid, Bases and Salts makes up 37.5% and Oxidation-Reduction makes up 25.0% of the examination. The cognitive level of the items varies from knowledge to understanding to higher mental processing. These are a simplification of the cognitive levels as described by Bloom's Taxonomy (1956). In the multiple choice component, the majority of the items are evaluating understanding. Lesser numbers of items evaluating knowledge and even fewer items evaluating higher mental processing (see Appendix B). The open-ended component focuses on understanding, but it may also include items of knowledge and higher mental processing chosen across the curricular areas.

Procedure

Students' response data from January 1999 and June 1999 were provided by the Ministry of Education. The personal information was collected in the Transcripts and Examinations (TRAX) System and Student Level Data System (SLD), in which student courses, percentages, provincial examination scores, and demographics are recorded. There was no disclosure of personal information and no attempt to link or match records. The Ministry of Education removed the names from the data files. Only the demographic information and response data were used for the analysis.

Model

The conceptual basis of the model used to analyze the data is that the relationship between item difficulty and student ability determines the performance of students on a test item. That is, a student with greater ability should also have a greater chance of success on a specific question than would a less able person. Conversely, a person of any level of ability would have a greater chance of success on a less-difficult question than on a more-difficult question. To the extent that this relationship holds, the probability of the success of a student on a question can be specified as a function of the difference between the ability of the student and the difficulty of the question.

The relationship between the examinee's item performance and the trait underlying the item performance can be described by a monotonically increasing function called an item characteristic function or item characteristic curve (Hambleton, Swaminathan & Rogers, 1991). An item characteristic curve is a mathematical expression that relates the probability of success on an item to the ability measured on the test and the characteristics of the item. Item characteristic curves for the one parameter logistic model (Rasch) are given by the equation:

$$P_i = e^{(B_n - D_i)} \qquad i = 1, 2, \ldots, n$$

where P_i is the probability of an examinee correctly answering item i

B_n is the proficiency level of examinee n, and

D_i is the difficulty level for item i

Note for the one parameter model, only the difficulty parameter, D_i exists.

Use of the Rasch Model involves the collection of the responses to a set of test items and an estimation of the values in the parameters in the model for the items and the students that best fit the data. Iterative computer procedures are used to calculate the maximum likelihood estimates of the parameters. Initial estimates are made for item difficulties based on the number of correct answers. Then initial estimates are made for the abilities of the students based on their scores. The initial estimates of ability are then used to improve the estimates of item difficulty, which in turn are used to improve the estimates of student ability. The process is iterated to maximize the fit of the parameter estimates to the test data.

Methods based on item response theory (IRT) provide a useful theoretical framework for DIF because between-group differences in the item parameters for the specific model can be used to model DIF. All the various DIF methods conceptualize DIF in terms of differences in the model parameters for the comparison groups (Clauser & Mazur, 1998). The general framework involves estimating item parameters separately for the groups studied.

Rasch DIF can be ascertained by examining the difficulty estimates. The estimates should be similar if the item displays no DIF. The amount of error in the estimates for both groups needs to be considered before deciding if the item truly does demonstrate DIF. An item would demonstrate DIF if the estimates from two different groups and their respective error do not overlap each other.

Difficulty estimates were produced using three Rasch programs: RASCAL version 3.51 (Assessment Systems Corporation, 1994), BIGSTEPS version 2.61 (Linacre & Wright, 1995) and FACETS version 3.04 (Linacre, 1996). The programs all use the same iterative process to arrive at the difficulty estimates. The intention was to establish that all three programs result in estimates for difficulty for the items regardless of the program used. The estimates were compared using a scatter plot with confidence intervals equivalent to the standard error of measurement. The estimates were found to be similar and within ± 1 standard error of estimate regardless of the program used.

Difficulty estimates for the various groups, e.g. male and female, were subjected to the Student's t-test to compare the similarity of the estimates. This served as a preliminary screening for DIF. Further examination of the difficulty estimates using protocol from Draba (1977) suggest that estimate shifts of one half of a logit or ten percent of the range would be conspicuous. Further analyses involved the use of scatter plots. The pairs of difficulty estimates are plotted with the identity line, x=y, added. Since the difficulty estimates are truly estimates, both X and Y confidence intervals equivalent to the standard error of measurement are placed around the point. If the confidence interval for an item overlaps the line of identity, the item is deemed to be free of DIF.

Polytomous DIF

The Rasch model can be extended for polytomous items. The open-ended component of the final examination can be analyzed using the partial credit model from the Rasch family of models that have been developed to deal with diverse response formats. A simple extension of right/wrong scoring allows the identification of one or more intermediate levels of performance on an item and the award of partial credit for reaching these intermediate levels (Wright & Masters, 1982).

A general expression for the probability of an event happening according to the partial credit model is as follows:

$$P_{nix} = \frac{e^{\sum_{j=0}^{x}(B_n - D_{ij})}}{\sum_{k=o}^{m_i} e^{\sum_{j=0}^{k}(B_n - D_{ij})}}$$

$x = 0, 1, \ldots, m_i$

where	P_{nix}	probability of person n scoring x on item i
	B_n	ability/attitude of person n
	D_{ij}	difficulty of step j in item i over step j - 1
	m_i	number of steps in item i

Item step parameters can be determined and consequently compared. The analysis was done using either BIGSTEPS version 2.61 (Linacre & Wright, 1995) and FACETS version 3.06 (Linacre, 1996). Differential item functioning can be ascertained at the various levels or for the entire item. These estimates were obtained for all of the comparison groups. The estimates were shown along with the error for each estimate. These estimates were presented as graphs showing the difficulty plotted against the item step. Evidence of DIF was considered to be present if the confidence levels did not overlap.

Chapter Three

Results

Before DIF analyses were performed, an examination of the scores was carried out to provide information about the performance of comparison groups over each of the subtests of the provincial examination. The differences in the mean scores are subject to the interpretation of the specific interest groups. An examination of only the means and standard deviations shows that the comparison groups exhibited different mean scores for the multiple-choice component and the open-ended component depending on the variable considered (see Table 1 & 2). Females were outscored by males. Students from large school districts outscored their counterparts from smaller districts. Non-aboriginals outscored aboriginals. These observations were seen in both sittings of the examination, which suggested that these results are consistent. These trends are subject to further investigation beyond comparing their means.

Insert Table 1

An examination of the descriptive statistics shows that the comparison groups exhibited different mean scores for the multiple-choice. Further examination of the mean scores to determine their statistical significance through t-testing showed that some of the differences are statistically significant. The small district versus large district was found to be different favoring large districts for both sittings; the January sitting ($t = 3.48$, $df = 2728$, $p < .001$) and June sitting ($t = 7.94$, $df = 7257$, $p < .001$). Similarly, the scores for non-aboriginals and aboriginals for January ($t = 2.42$, $df = 3507$, $p < .05$) and June ($t = 3.66$, $df = 8956$, $p < .001$) were also found to be statistically significant. The June results comparing school districts showed differences at various alpha levels. Small district versus medium district means were found to be different ($t = 2.68$, $df = 3423$, $p < .01$). The same was true for medium district versus large district means ($t = 4.55$, $df = 7152$, $p < .001$). Cohen's d

Table 1

Mean Scores for the Multiple Choice Component

	January				June			
Group	n	M	(SD)	Cohen's d	n	M	(SD)	Cohen's d
Females	1746	31.0	(8.0)	0.04	4506	**30.9**	**(7.8)**	0.09
Males	1763	31.3	(8.5)		4452	**31.6**	**(8.3)**	
Small District	1073	**30.7**	**(7.9)**	0.12	1759	30.1	(7.9)	0.00
Medium District	631	**29.8**	**(7.9)**	0.26	1097	**30.1**	**(7.8)**	0.21
Large District	1805	**31.9**	**(8.5)**	0.14	6102	**31.8**	**(8.1)**	0.21
Non-aboriginals	3471	**31.2**	**(8.2)**	0.47	8883	**31.3**	**(8.1)**	0.64
Aboriginals	38	**27.3**	**(7.0)**		75	**26.1**	**(6.9)**	
Province	3509	31.1	(8.2)		8958	31.2	(8.1)	

Note. Maximum score = 48. n= number of students; M= mean; SD= standard deviation.

The bolded means show significant differences from t-testing ($\alpha = 0.05$). Effect sizes for the school districts show the comparison between small versus medium; medium versus large, and large versus small with the first district listed as the reference cell.

effect sizes were calculated in every instance from the means and the standard deviation of the general population. These were found to range from 0.67 to 0.04. Interpretation of Cohen's d tells us that the distribution of scores overlap for both groups and the results of the t-test are not of practical significance. The effect sizes were all considered trivial with the exception of the effect size for the nonaboriginal - aboriginal difference, which is classified as large.

A t-test analysis of the open-ended component showed that there were also statistically significant differences for this component. The female versus male comparison was found to be different for the January sitting ($t = 2.40$, $df = 3507$, $p < .05$). The small district versus medium district comparison was also found to be significantly different. The difference favored small districts during the January sitting ($t = 2.64$, $df = 1846$, $p < .01$) and favoring the medium district during the June sitting ($t = 4.44$, $df = 3493$, $p < .01$). The medium versus large district comparison was found to be significantly different and it favored the large district for both the January sitting ($t = 3.93$, $df = 2435$, $p < .01$) and the June sitting ($t = 4.54$, $df = 7199$, $p < .01$). For the June sitting, mean scores for the large and small districts differed significantly ($t = 10.0$, $df = 7224$, $p < .001$). In both sittings, the non-aboriginals versus aboriginal means were also found to be significantly different during the January sitting ($t = 2.57$, $df = 3507$, $p < .01$) and June sitting ($t = 6.04$, $df = 8958$, $p < .001$). Once again, effect sizes were calculated and the results were all considered trivial with the exception of the variable of aboriginal status. This variable had a large effect size.

Insert Table 2

From the mean scores for each of the subtests, one may speculate that certain results are consistent with specific patterns that have been observed over many testing periods. The

Table 2

Mean Scores for the Open Ended Component

	January				June			
Group	n	M	SD	Cohen's d	n	M	SD	Cohen's d
Females	1746	**20.9**	**7.1**	0.08	4506	18.4	7.7	0.03
Males	1763	**20.3**	**7.7**		4452	18.6	8.2	
Small District	1074	**20.6**	**7.1**	0.12	1759	**16.9**	**8.0**	0.15
Medium District	772	**19.7**	**7.3**	0.17	1734	**18.1**	**7.9**	0.13
Large District	1663	**21.0**	**7.7**	0.05	5465	**19.1**	**7.9**	0.28
Non-aborignal	3471	**20.6**	**7.4**	0.43	8883	**18.5**	**7.9**	0.72
Aboriginal	38	**17.5**	**7.1**		75	**13.0**	**7.4**	
Province	3509	20.6	7.4		8958	18.5	8.0	

Note. Maximum score = 32. n= number of students; M= mean; SD= standard deviation.

The bolded means show significant differences from t-testing. Effect sizes for the school districts show the comparison between small versus medium; medium versus large, and large versus small with the first district listed as the reference cell.

differences may be used to initiate change from a pedagogical standpoint. Other differences may be too difficult to reconcile because of a diverse set of circumstances.

Differential Item Functioning

The comparison of means, whether by statistical testing or calculation of effect size indices, was used to establish the relative performance of the subgroups that wrote the Chemistry 12 examinations. In any item response model, the Rasch model included, these differences in proficiency for individuals within groups must be conditioned out mathematically in order to produce "sample free" difficulty estimates of the items. The effect of DIF is said to exist when an item has statistically significant differences in difficulty estimates.

A comparison of means is not sufficient in the analysis of the subtest or items to see if that same subtest or items adequately discriminate among the stakeholders. Consideration of the means does not take into consideration the composition of the group with respect to ability. Looking for DIF removes group composition and shifts the emphasis to the performance of the individual items after equating for ability. Only after this is done, can the items that comprise a test be used to obtain valid inferences and eliminate other interpretations.

Rasch item difficulty estimates and the associated standard errors were calculated for each of the elements within the three independent variables of interest. The BIGSTEPS software was used to determine the difficulty estimates. The results of this analysis are displayed in Table 3 for the January 1999 multiple-choice component for the first twelve items. The other difficulty estimates are included in Appendix C. The difficulty estimates needed to be further scrutinized to verify if DIF does exist in the item's performance.

From the estimates, one can see that the majority of the items have similar estimates when one considers the standard error of measurement. A few items appear troublesome using the methodology of Draba (1977). These items differ by more than half a logit or ten percent of the range of the estimates.

Insert table here (Table 3)

Item 4 from the January multiple-choice component performs well when considering the standard error of the estimate across the comparison groups. When considering gender, the difficulty estimate from the female group was 0.98 (0.05) and the male difficulty estimate was 1.03 (0.05). The difficulty estimates would be considered identical. When considering district size, the small districts' difficulty estimate was 1.00 (0.07), the medium districts' difficulty estimate was 1.08 (0.08) and the large districts' difficulty estimate was 0.97 (0.06). Again considering the error, the estimates are identical. The last variable, of aboriginal status, resulted in difficulty estimates of 1.00 (0.04) for non-aboriginals and 1.05 (0.53) for aboriginals. Note the large error associated with this estimate as a consequence of small sample size. This item did not display evidence of DIF for any of the three variables.

Item 7, of the same sitting, did not perform well. The gender comparison resulted in difficulty estimates of -0.22 (0.06) for females and -0.62 (0.06) for males. The district size comparison resulted in difficulty estimates of -0.16 (0.07) for small districts, -0.34 (0.09) for medium districts and -0.63 (0.07) for large districts. The aboriginal status comparison resulted in difficulty estimates of -0.41 (0.04) for non-aboriginals and -1.24 (0.65) for aboriginals. This item showed DIF across all three comparisons. This item was the only item that showed significant differences for our three comparisons; other items may have showed significant differences for two groupings but the majority showed a difference only for one comparison.

For all the pairs of difficulty estimates for our comparison groups, t-testing was performed as a method to screen the items. This testing, even though cumbersome, did identify certain items as having statistically significant difficulty measures. Further, the methodology of Draba (1977) was utilized to identify items with significant differences in difficulty estimates. Looking for differences greater than 0.5 logits or ten percent difference

Table 3

Difficulty Estimates for January 1999 Multiple Choice Component (Items 1 - 12)

Item	Gender				School District Size						Aboriginality			
	Female		Male		Small		Medium		Large		Non-aboriginals		Aboriginals	
	D_i	SE	D_i	SE	D_i	SE	D_i	SE	D_i	SE	D_i	SE	D_i	SE
1	0.22	(0.05)	-0.06	(0.06)	0.29	(0.07)	0.04	(0.08)	-0.04	(0.06)	0.08	(0.04)	-0.26	(0.53)
2	-1.80	(0.09)	-2.30	(0.10)	-2.09	(0.12)	-2.55	(0.16)	-0.79	(0.09)	-2.02	(0.07)	deleted	deleted
3	-0.51	(0.06)	-0.34	(0.06)	-0.65	(0.08)	-0.36	(0.09)	-0.31	(0.06)	-0.43	(0.04)	0.01	(0.52)
4	0.98	(0.05)	1.03	(0.05)	1.00	(0.07)	1.08	(0.08)	0.97	(0.06)	1.00	(0.04)	1.05	(0.53)
5	-1.57	(0.08)	-1.66	(0.08)	-1.48	(0.10)	-1.76	(0.12)	-1.64	(0.09)	-1.62	(0.06)	-1.72	(0.76)
6	-1.18	(0.07)	-1.13	(0.07)	-1.17	(0.09)	-1.20	(0.10)	-1.12	(0.07)	-1.15	(0.05)	-2.48	(1.02)
7	-0.22	(0.06)	-0.62	(0.06)	-0.16	(0.07)	-0.34	(0.09)	-0.63	(0.07)	-0.41	(0.04)	-1.24	(0.65)
8	-0.56	(0.06)	-0.60	(0.06)	-0.50	(0.08)	-0.49	(0.09)	-0.68	(0.07)	-0.58	(0.04)	-0.86	(0.59)
9	0.04	(0.05)	-0.12	(0.06)	-0.05	(0.07)	0.06	(0.08)	-0.08	(0.06)	-0.04	(0.04)	-0.26	(0.53)
10	2.44	(0.07)	2.52	(0.07)	2.43	(0.08)	2.52	(0.10)	2.50	(0.07)	2.49	(0.05)	1.34	(0.56)
11	-0.33	(0.06)	-0.33	(0.06)	-0.43	(0.08)	-0.39	(0.09)	-0.23	(0.06)	-0.33	(0.04)	-0.54	(0.55)
12	0.30	(0.05)	0.11	(0.06)	0.22	(0.07)	0.25	(0.08)	0.18	(0.06)	0.21	(0.04)	0.52	(0.51)
n	1744		1760		1071		773		1059		3492		17	
df	19		19		19		19		19		19		3	

Note. Item 2 did not generate difficulty estimates; the response data was not consistent with the Rasch model.

is much easier and provided similar information. Both of these methods identified items that displayed DIF that were identified using our third method. Ultimately from the scatterplots of the difficulty estimates, the items were identified as displaying DIF; therefore tallies were made to see which group was advantaged.

The results of the DIF analysis for the three factors of study showed that DIF was present in some of the items that comprised the multiple-choice components of both sittings. Depending on the reference and comparison groups, DIF was found in varying degrees (see Table 4). By comparing the difficulty estimates along with their errors of estimate, DIF was found in as few as fourteen items in the June 1999 sitting for the aboriginality grouping and for as many as thirty items for the June 1999 sitting of the examinations between small and large school districts. Differential item functioning was found to occur in all comparison groups. In many instances, the DIF effects negated one another. Other comparisons had high instances of DIF that did not negate each other. For these comparisons one group was favored over the other.

Both the January 1999 and June 1999 multiple-choice component subtests exhibited items that displayed DIF. The results are displayed in Table 4. The January 1999 multiple-choice subtest showed many instances of DIF. With respect to gender, ten items favored the females while ten items favored the males. As a result, there was no net DIF. When looking at district size as a variable, small districts were favored on nine items while medium districts were favored on twelve items. Overall, three items favored the medium-sized school district. Small districts were favored on ten items while large districts were favored on eleven items; large school districts had the advantage. The last district favored the medium districts on eleven items while large districts were favored on nine items. Overall, the medium districts had advantage on two items. The comparison between non-aboriginals and aboriginals had an equal number of items showing DIF so the DIF was negated. Overall, net DIF was not noticeable in the forty-eight item subtest.

Insert Table 4

The June 1999 multiple-choice component subtest was examined in the same manner. For the gender variable, thirteen items favored the males where sixteen items favored the females. Small districts were favored on eleven items while medium districts were favored on five items. Small districts were favored on fourteen items while large districts were favored on sixteen items. Medium districts were favored on twelve items and large districts were favored on thirteen items. Aboriginals were favored on six items while their non-aboriginal counterparts were favored on eight items. Once again the net effects of DIF were minimal for each of our comparisons.

For ease of analysis, scatter plots of the difficulty estimates were used to enumerate the items that showed DIF. The primary function of this analysis was to check to see if the items displayed DIF. This approach can be used to examine the quality of the subtest. The number of items and their position relative to the line of identity can be used for the two comparison groups. Those items above line favored the group found on the ordinate axis while those items found below the line favored the group found on the abscissa axis. Items found straddling the line of identity were judged not to display any DIF. Those items deviating from the line of identity can be identified from the graphs rather than from the tables. This method provides a quick method of determining the number of items that favor each group (see Figure 1). While the identity of the item cannot be discerned easily from the graph, it can be determined from its coordinates. Equivalent numbers of items were found to show DIF using either the graph and the examination of difficulty estimates using t-testing or the logit difference.

The sensitivity of this method can be seen to be dependent on the magnitude of the error of estimate for difficulty. Error bars coincide with the error of estimates for each of the comparison groups. The error bars running horizontally are for the reference group while the error bars running vertically are for the comparison group. The size of the error

Table 4

Summary of the Occurrence of DIF In the Multiple Choice Component

Variable	January					June				
Gender	Females	10	vs	Males	10	Females	16	vs	Males	13
District Size	Small	9	vs	Medium	12	Small	11	vs	Medium	5
	Small	10	vs	Large	11	Small	14	vs	Large	16
	Medium	11	vs	Large	9	Medium	12	vs	Large	13
Aboriginal Status	Non-aboriginals	9	vs	Aboriginals	9	Non-aboriginals	8	vs	Aboriginals	6

<u>Note</u>. There are 48 items that comprise the multiple-choice subtest. The number of occurences of DIF is shown for each comparison.

bars, especially when considering the error in estimate for the aboriginal group, shows a lack of sensitivity when the number of respondents in the group is small. Excluding the aboriginal group, the other groups were large enough to generate stable estimates of difficulty with small degrees of error

Insert Figure 1.

Consistency of DIF Across Sittings

Examining the differential item functioning over the two sittings for the gender variable illustrates the occurrences of DIF. The results for the first twenty items that cover three curricular areas are shown (see Table 5). The examinations are constructed from a table of specifications. The DIF that is replicated for specific enumerated items only shares similar general curricular area. Item 1 and 2 of the January sitting favors males. Item 11 favored females in the June sitting. Item 3 favored females during the January sitting and favored males during the June sitting. These items possess dissimilar learning outcomes. For both sittings, item 7 favored the males, while item 15, favored the females. The latter differences occurred for similar but not identical learning outcomes. Overall consideration of the entire multiple-choice sub-test revealed four instances of DIF that replicated for similar items for females and three instances of DIF that were replicated for similar items for males. Similar findings existed for the other variables. Note that the DIF is found only on identically numbered items that share common curricular content but not always at the level of the prescribed learning outcome.

Insert Table 5

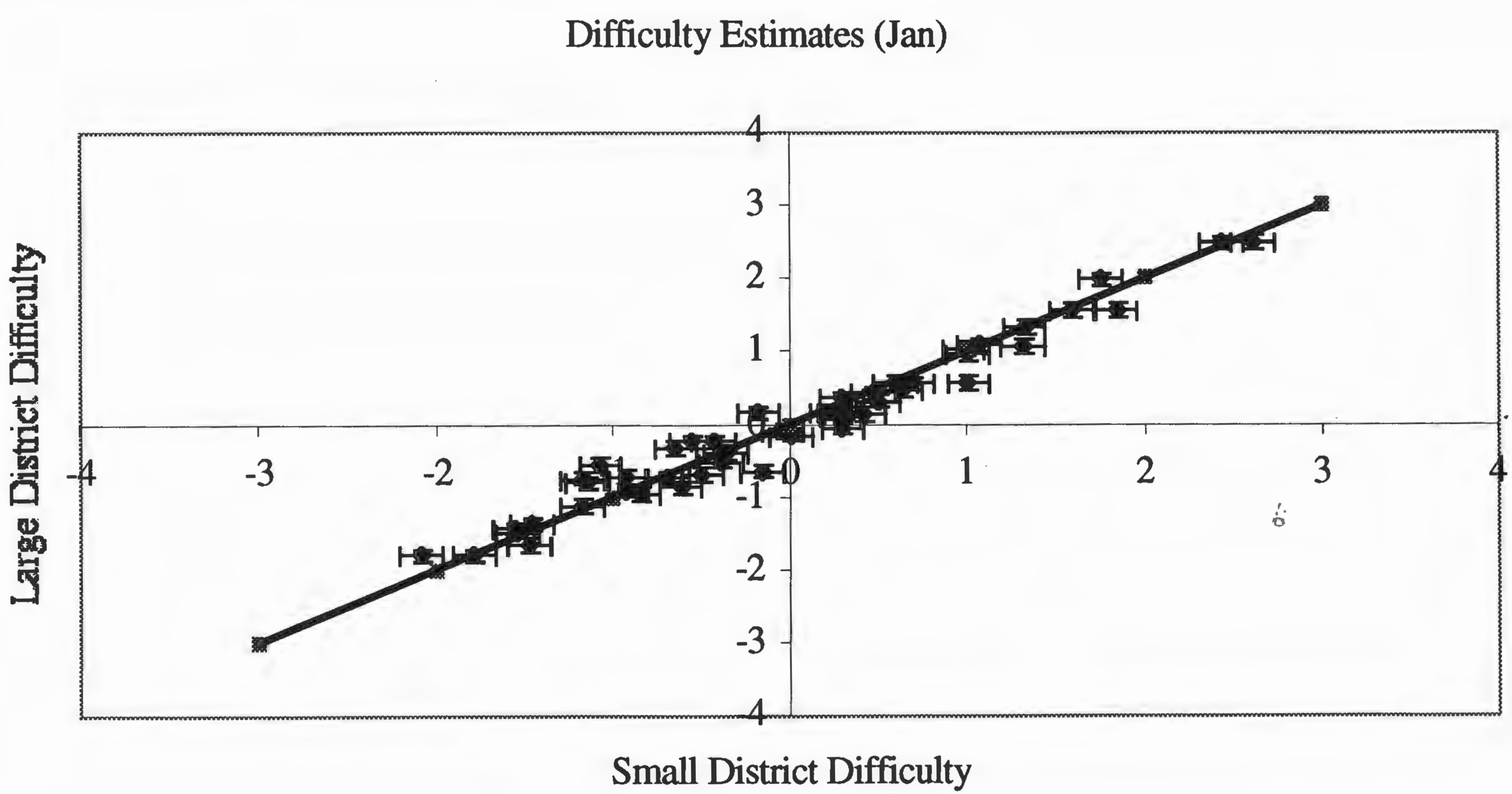

Figure 1. Difficulty estimates for January 1999 multiple-choice items from small and large school districts. Items found overlapping the line of identity do not display DIF.

Table 5

The Occurrence of Differential Item Functioning With Respect to Prescribed Learning Outcome (PLO) For Items 1 - 17.

Item	Curriculum Organizer	Jan Female	Male	June Female	Male	January PLO	June PLO
1	1		Male			A3	A2
2	1		Male		Male	B3	A2
3	1	Female			Male	B2	A6
4	1				Male	B4	B6
5	1				Male	B6	B3,9
6	1					C5	C3
7	2		**Male**		**Male**	**D4**	**D7**
8	2					E2	E2
9	2					E2,5	E2,5
10	2		Male			E3	E4
11	2			Female		F3	F2
12	2					F4	F4
13	3	**Female**		**Female**		**F7**	**F7**
14	3					G4	G8
15	3	**Female**		**Female**		**H2**	**H1**
16	3			Female		H3	H7
17	3	**Female**		**Female**		**I2**	**I3**

Note. The bolded items show replication of DIF within the same general curriculum area. Curriculum organizers are as follows: 1 = Chemical Kinetics; 2 = Chemical Equilibrium; 3 = Chemical Solubility. The PLO's are further divisions of the general curricular areas.

Cursory examination of the occurrences of DIF would show that the majority of items that did show DIF are not correlated when considering the parallel nature of the examinations. The DIF was not repeated for specific items of curricular content or for the students' cognitive levels. Each examination is unique. Items of similar curricular content behaved differently from examination to examination. Each item has a prescribed learning outcome. The items are similar only by their enumeration, but there are instances where the same numbered item has the same prescribed learning outcome. Item 13 had the same prescribed learning outcome on both forms and the DIF favored the females over the two sittings. Item 40 also had the same prescribed learning outcome and it too had DIF that favored females. Item 12 had the same prescribed learning outcome but it did not display any DIF.

Items may have shown DIF for the same group but it should be noted that the specific learning outcome usually is different even though it falls under the same general curriculum organizer; for example, in both sittings item 2 favored males but the prescribed learning outcome was different. Assessment of the examination specifications showed that the prescribed learning outcomes covered by the multiple-choice components were genuinely unique for each subtest. The majority of the items did not show DIF. The specific items demonstrating DIF are known and further analyses is possible if there is a need to see if the general curriculum area is found easier for one group or the other.

The open-ended component of the provincial examination was subject to the same DIF analysis based on the overall difficulty estimate for the items. Since the open-ended component consisted of items that merited partial marks, further analysis was performed to see if DIF was found over the range of mark values for each item.

Difficulty estimates along with the error in estimates were generated using BIGSTEPS (Linacre & Wright, 1995) software. These estimates were compared across the comparison groups. Each item was examined to see if it displayed DIF and then the net effects of DIF were assessed over the subtest. Comparisons for each subtest were made between the January and June sitting to see if DIF was found and replicated.

The January 1999 open-ended component subtest showed differential item or the three factors of this study (see Table 6). The same criteria was used for screening items for DIF and then scatterplot analysis was subsequently utilized. When it came to gender, males were favored on one of the items while females were favored on two items. The factor school district size showed that the smaller districts were disadvantaged by one item versus their medium district counterparts and favored by two items in comparison to the large school districts. The medium school district was favored on five items relative to large districts while the latter were favored over the former on three items. The third factor, aboriginal status, had aboriginals favored on one item.

Insert table 6 here.

The June 1999 open-ended component subtest also showed DIF occurring for each of the respective factors (see Table 7). With respect to gender, males were favored for three items while females were favored on six items. The small district comparison showed no net DIF since the number of items showing DIF in either comparison was equal. Medium districts were favored on five items while their comparison group of large districts were favored on four items. The third variable, aboriginal status, showed that both groups were favored on four items so no net DIF was observed.

Insert table 7 here.

Difficulty estimates were again screened using t-testing and the criterion of greater than 0.5 logits difference in difficulty estimates. The items were flagged and subjected to further comparison. These estimates were plotted against the line of identity. The plotting of the difficulty estimates provides a convenient means of looking for DIF for each item as

Table 6

Difficulty Estimates for January 1999 Open-Ended Component

Item	Gender				District Size						Aboriginal Status			
	Female		Male		Small		Medium		Large		Non-aboriginal		Aboriginal	
	D_i	(SE)	D_i	(SE)	D_i	(SE)	D_i	(SE)	D_i	(SE)	D_i	(SE)	D_i	(SE)
1	0.28	(0.02)	0.32	(0.02)	**0.38**	**(0.03)**	**0.23**	**(0.03)**	**0.33**	**(0.03)**	0.30	(0.02)	0.17	(0.15)
2	**-0.78**	**(0.03)**	**-1.02**	**(0.03)**	**-0.84**	**(0.04)**	**-0.69**	**(0.03)**	**-1.17**	**(0.03)**	-0.90	(0.02)	-0.77	(0.20)
3	-0.35	(0.03)	-0.32	(0.03)	**-0.32**	**(0.04)**	**-0.27**	**(0.03)**	**-0.40**	**(0.03)**	-0.33	(0.02)	-0.20	(0.17)
4	-0.08	(0.03)	-0.05	(0.03)	-0.07	(0.03)	**0.01**	**(0.03)**	**-0.12**	**(0.03)**	-0.06	(0.02)	-0.08	(0.15)
5	**-0.48**	**(0.03)**	**-0.40**	**(0.03)**	-0.46	(0.04)	-0.41	(0.03)	-0.45	(0.03)	-0.44	(0.02)	-0.28	(0.16)
6	0.35	(0.02)	0.35	(0.02)	**0.31**	**(0.03)**	**0.33**	**(0.03)**	**0.40**	**(0.03)**	0.35	(0.02)	0.30	(0.15)
7	**0.01**	**(0.02)**	**0.11**	**(0.02)**	**0.04**	**(0.03)**	**-0.04**	**(0.03)**	**0.17**	**(0.03)**	0.06	(0.02)	0.01	(0.15)
8	0.19	(0.03)	0.20	(0.03)	**0.19**	**(0.03)**	**0.11**	**(0.03)**	**0.28**	**(0.03)**	0.19	(0.02)	0.32	(0.16)
9	0.85	(0.03)	0.80	(0.03)	**0.77**	**(0.03)**	**0.74**	**(0.03)**	**0.96**	**(0.03)**	**0.83**	**(0.02)**	**0.53**	**(0.17)**
n	1744		1760		1071		773		1059		3492		17	
df	8		8		8		8		8		8		8	

Note. The bolded difficulty estimates show items that show differential item functioning. The difficulty estimates considered with the standard errors of estimates do not overlap.

Table 7

Difficulty Estimates for the June 1999 Open-Ended Component

Item	Gender		School District Size			Aboriginal Status	
	Female	Male	Small	Medium	Large	Non-aboriginal	Aboriginal
	D_i (SE)	D_i (SE)	D_i (SE)	D_i (SE)	D_i (SE)	D_i (SE)	D_i (SE)
1	**0.18 (0.01)**	**0.11 (0.01)**	0.14 (0.02)	**0.12 (0.02)**	**0.16 (0.01)**	**0.15 (0.01)**	**-0.03 (0.08)**
2	**-0.45 (0.01)**	**-0.59 (0.01)**	**-0.47 (0.01)**	**-0.48 (0.02)**	**-0.56 (0.01)**	**-0.52 (0.01)**	**-0.34 (0.07)**
3	**0.13 (0.01)**	**0.22 (0.01)**	**0.10 (0.02)**	**0.17 (0.02)**	**0.21 (0.01)**	**0.18 (0.01)**	**-0.16 (0.07)**
4	**0.13 (0.01)**	**0.19 (0.01)**	**0.12 (0.02)**	**0.14 (0.02)**	**0.19 (0.01)**	**0.16 (0.01)**	**0.00 (0.08)**
5	**-0.23 (0.01)**	**-0.38 (0.01)**	**-0.25 (0.02)**	**-0.22 (0.02)**	**-0.35 (0.01)**	**-0.30 (0.01)**	**-0.07 (0.08)**
6	0.99 (0.02)	0.97 (0.01)	**0.90 (0.02)**	**0.93 (0.02)**	**1.03 (0.01)**	0.98 (0.01)	0.93 (0.13)
7	**-0.51 (0.01)**	**-0.46 (0.01)**	**-0.39 (0.02)**	**-0.45 (0.02)**	**-0.54 (0.01)**	**-0.49 (0.01)**	**-0.32 (0.07)**
8	**-0.35 (0.01)**	**-0.30 (0.01)**	**-0.29 (0.02)**	**-0.30 (0.02)**	**-0.35 (0.01)**	-0.33 (0.01)	-0.32 (0.07)
9	-0.68 (0.01)	-0.66 (0.01)	**-0.60 (0.02)**	**-0.70 (0.02)**	**-0.69 (0.01)**	**-0.67 (0.01)**	**-0.53 (0.07)**
10	**0.07 (0.01)**	**0.12 (0.01)**	**0.02 (0.02)**	**0.06 (0.02)**	**0.13 (0.01)**	**0.09 (0.01)**	**-0.05 (0.07)**
11	**0.73 (0.01)**	**0.78 (0.01)**	**0.70 (0.02)**	**0.75 (0.02)**	**0.78 (0.01)**	0.75 (0.01)	0.88 (0.12)
n	4490	4429	1765	1660	5494	8923	35
df	10	10	10	10	10	10	10

Note. The bolded difficulty estimates show items that show differential item functioning. The difficulty estimates considered with the standard errors of estimates do not overlap.

well as a general idea of the net DIF that is exhibited over the subtest (see Figure 2). The confidence intervals are very small due to the large numbers in the majority of the comparison groups. The exception, once again, are the confidence intervals for the aboriginals. Their small sample generated unstable estimates for difficulty. With these small errors in estimate, the sensitivity of the plot is much more heightened and the value of plots provides relief from the examination of estimates from tables. However, with these small differences, perhaps meaningless differences are declared as statistically significant. Once again, the effect size should be taken into consideration. The confidence intervals can be increased to multiples of the standard error to reduce the number of items that exhibit DIF.

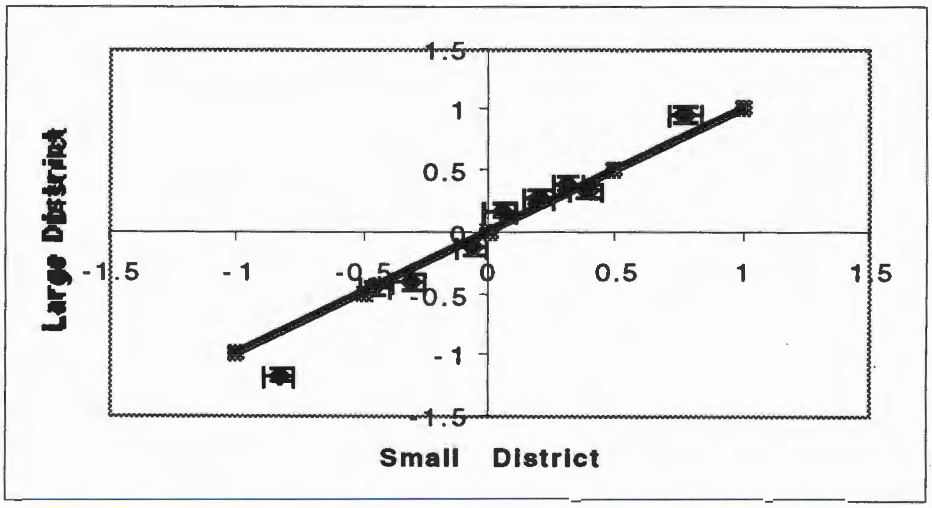

Figure 2. Difficulty estimates for the January 1999 open ended items from small and large school districts. The majority of items straddle the line of identity. One item is above the line and another item is below the line. The DIF is negated on the subtest.

The graphs were examined and the items were categorized as showing DIF or not showing DIF. Those items showing DIF were then categorized according to the comparison group that the item favored. These results are summarized in Table 8.

Insert Table 8

Looking at the occurrence of DIF from the two sittings showed that DIF was either replicated or DIF was isolated. If DIF was isolated, the reference group showed DIF in one sitting and the comparison showed DIF in the other sitting. With respect to gender, females were advantaged during both sittings. The advantage and disadvantage the small district had relative to their larger counterparts were not seen in the June open-ended component. DIF was replicated for the medium districts over the larger districts in this sitting. Considering the aboriginal status variable, the DIF was not replicated. For other examinations to be validated, these observations need to be further substantiated by examining the net DIF.

Examination of the open-ended subtest for each variable showed the overall effects of DIF favoring one group over another. Further examination of each item that comprised the open-ended subtest show that the DIF is repeated for items similar in enumeration, but not replicated for items involving the same prescribed learning outcomes. The composition of the open ended component makes each sitting unique. The number of items and the specific prescribed learning outcomes that are evaluated differ from sitting to sitting. There are instances where DIF is indeed replicated for similar items with the same prescribed learning outcomes, so those items should be subject to further analyses (see Table 9). Consider item 2. Males were favored over both sittings and the prescribed learning outcomes are not identical even though they belong to the same general curriculum content area. Looking at different items from the two different subtests that had the same curricular organizer, and, more importantly, similar prescribed learning outcomes, both item 8 from January and item 9 from June showed that DIF was absent. Valuable information can be extracted from different difficulty estimates concerning the performance of the group. DIF did occur for similarly enumerated items but different prescribed learning outcomes. However, these differences were not considered to be important for this investigation.

Table 8

Summary of the Occurrence of DIF in the Open Ended Component

Variable	January				June			
Gender	Females	2	Males	1	Females	6	Males	3
District Size	Small	1	Medium	2	Small	2	Medium	2
	Small	4	Large	2	Small	5	Large	5
	Medium	5	Large	3	Medium	5	Large	4
Aboriginal Status	Non-aboriginal	0	Aboriginal	1	Non-aboriginal	4	Aboriginal	4

Note. The occurence of DIF for each sitting is displayed above. Net DIF can be determined from the comparison of the two groups.

Insert Table 9

Partial Credit Analysis Of Open Ended Subtest

To further explore the existence of DIF in the items that comprise the open-ended component, partial credit analysis was performed. Difficulty estimates were obtained for every half mark increment for these items using BIGSTEPS. The difficulty estimates showed how the two comparison groups fared in obtaining marks for each of the items at the various mark values (see Table 10). The difficulty estimates can also be used to confirm the rubric set up to award part marks for the marking process.

Most of the items subject to this analysis showed that the estimates within experimental error were the same for each increment no matter what item was chosen and what factor. These estimates further showed that the scoring of students was not influenced by the factors considered during study. At two levels, both macroscopically and microscopically, items have been analyzed to show that they are DIF free.

An example of a DIF-free item would be item eight of the January sitting. The estimates and their errors for the half mark increments for this three mark item were obtained for the gender variable. The half mark estimates for females in increasing order are : -0.43 (0.12), -0.37 (0.10), 0.00 (0.10), 0.21 (0.09), 0.72 (0.08) and 0.73 (0.08). The half mark estimates for males in increasing order are: -0.42 (0.11), -0.20 (0.09), -0.01 (0.09) 0.23 (0.09), 0.67 (0.08) and 0.72 (0.07) (see Figure 3). When one considers the error estimates for only a single difficulty estimate, all of the values fall within the error bars except for the estimate for the score of one. For that score on this item, both error estimates need to be considered to see the overlap.

Insert Table 10

Table 9

Occurrence Of DIF in the Open Ended Component With Respect to Gender

January				June			
Item	Gender	CO	PLO	Item	Gender	CO	PLO
1		1	C2	1	Males	1	B9
2	**Males**	**2**	**D4; F1,5**	**2**	**Males**	**2**	**D3,4;F5**
3		3	G5, I3	3	Females	2	E2
4		4	M4, N1,3	4	Females	3	H3
5	Females	4	M3	5	Males	3	I4
6		4	P1, 4, 6	6		4	K7
7	Females	5	T2	7	Females	4	M3,4,5
8		5	T6	8	Females	4	P3
9		5	W6	9		5	T6
				10	Females	5	U1,7
				11	Females	5	W4
Total	Females	2		Total	Females	6	
DIF	Males	1		DIF	Males	3	

Note. Bolded item shows DIF for similar item examining similar prescribed learning outcomes. Curriculum organizers (CO) are as follows : 1 = Chemical Kinetics; 2 = Chemical Equilibrium; 3 = Solubility. Prescribed learning outcomes (PLO) are specific curricular content areas.

Table 10

January Partial Credit By Gender

Item	Gender	Partial Credit Step Difficulty Estimate																
		0.00	0.50		1.00		1.50		2.00		2.50		3.00		3.50		4.00	
		Step	Step	Error	Step	Error	Step	Error	Step	Error	Step	Error	Step	Error	Step	Error	Step	Error
3	Female	-1.13	-0.64	(0.14)	-0.48	(0.12)	-0.30	(0.10)	-0.01	(0.08)	0.19	(0.07)	0.48	(0.07)	0.85	(0.07)	1.36	(0.07)
	Male	-1.24	-0.73	(0.16)	-0.65	(0.13)	-0.41	(0.10)	-0.08	(0.08)	0.15	(0.08)	0.58	(0.07)	0.82	(0.07)	1.37	(0.07)
4	Female	-0.98	-0.66	(0.15)	-0.40	(0.10)	-0.13	(0.08)	0.11	(0.07)	0.32	(0.07)	0.68	(0.07)	1.09	(0.07)	1.73	(0.08)
	Male	-0.93	-0.70	(0.12)	-0.70	(0.12)	-0.36	(0.09)	-0.11	(0.08)	0.06	(0.08)	0.45	(0.07)	1.24	(0.07)	1.85	(0.08)
5	Female	-1.24	-0.83	(0.18)	-0.39	(0.11)	-0.20	(0.09)	-0.22	(0.09)	0.03	(0.08)	0.27	(0.08)	0.65	(0.07)	1.14	(0.06)
	Male	-1.09	-0.93	(0.15)	-0.51	(0.10)	-0.26	(0.09)	0.04	(0.09)	0.04	(0.09)	0.32	(0.08)	0.65	(0.07)	1.25	(0.06)
6	Female	-0.88	-0.30	(0.10)	-0.03	(0.08)	0.23	(0.07)	0.47	(0.07)	0.86	(0.07)	1.34	(0.08)	1.87	(0.10)	2.07	(0.15)
	Male	-0.84	-0.47	(0.10)	-0.13	(0.08)	0.15	(0.07)	0.46	(0.07)	0.84	(0.07)	1.30	(0.08)	1.95	(0.10)	2.17	(0.14)
7	Female	-0.38	0.18	(0.11)	0.00	(0.08)	0.21	(0.08)	0.57	(0.08)	1.09	(0.08)	1.11	(0.09)				
	Male	-0.30	0.15	(0.11)	-0.07	(0.08)	0.14	(0.08)	0.43	(0.08)	1.12	(0.08)	0.90	(0.10)				
8	Female	-0.68	0.10	(0.12)	-0.37	(0.10)	0.00	(0.10)	0.21	(0.09)	0.72	(0.08)	0.73	(0.08)				
	Male	-0.74	-0.42	(0.11)	-0.20	(0.09)	-0.01	(0.09)	0.23	(0.09)	0.67	(0.08)	0.72	(0.07)				

Note. Estimates are not available for items 1 and 2.

Further analysis of the open ended sub-test items using the partial credit model further strengthened the claim of the quality of the items. The majority of the questions from both sittings did not show any DIF. Examination of their difficulty estimates showed that the estimates were within the error of estimation. This is apparent from examining the graphs of the difficulty estimates for each increment for the reference and comparison groups. The area between the two partial credit curves is minute. Those items subject to testing met this standard, which reinforced the assumed quality of the testing procedure and the marking process.

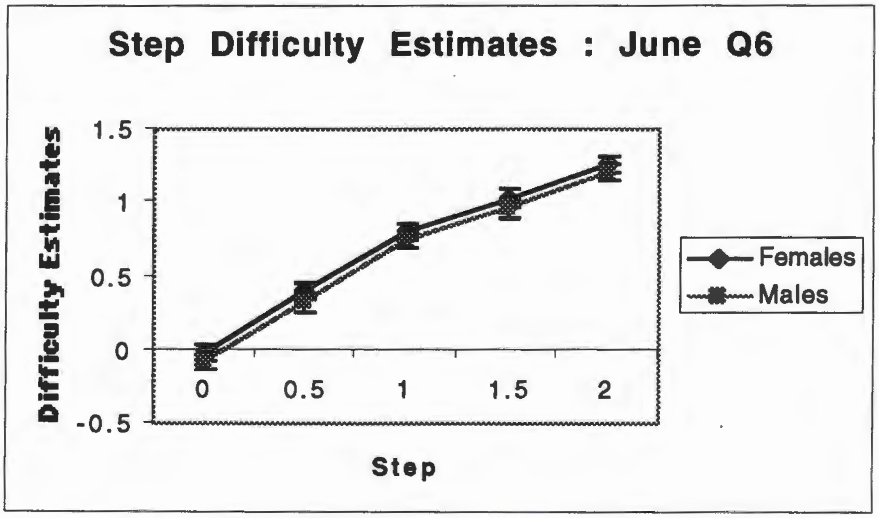

Figure 3. Difficulty estimates for the increments.

Difficulty estimates for this item shows that within the error of estimation, the estimates are the same.

Very few of the items subject to partial credit analyses demonstrated differential item functioning for any of the incremental scores within the items. None of the items showed DIF for all of the increments. If an item showed DIF, it was limited to less than one half of the increments and it showed DIF for the same group. There were some interesting patterns of differential item functions within specific items. DIF was found in one instance where at one increment the reference group was favored and at a higher increment, the comparison group was favored (see Figure 4). The DIF was isolated and showed no patterns or replication for any of the other variables. DIF was found in another instance at lower increments but DIF was not observed at higher increments (see Figure 5). Again this was an isolated incidence of DIF.

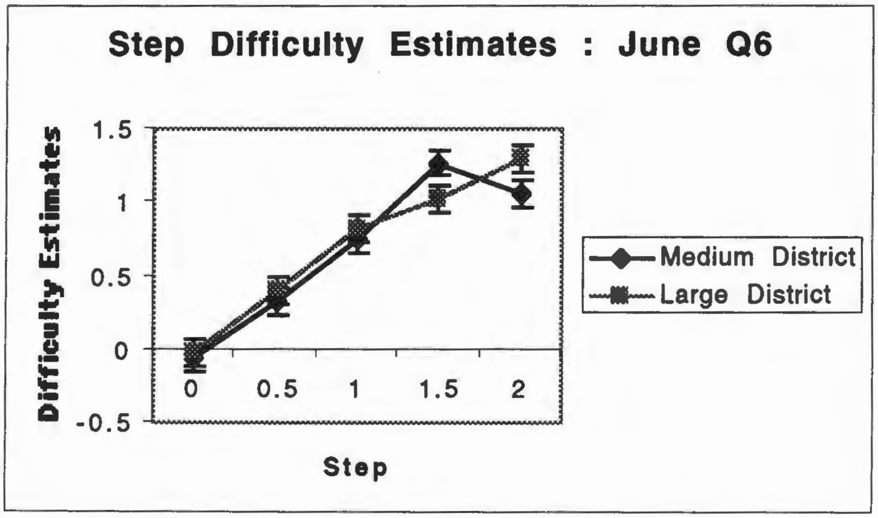

Figure 4. Difficulty estimates for the increments.

Difficulty estimates for this item showed DIF for both groups at different increments for the same question. Note the area between the two partial credit curves bulges out.

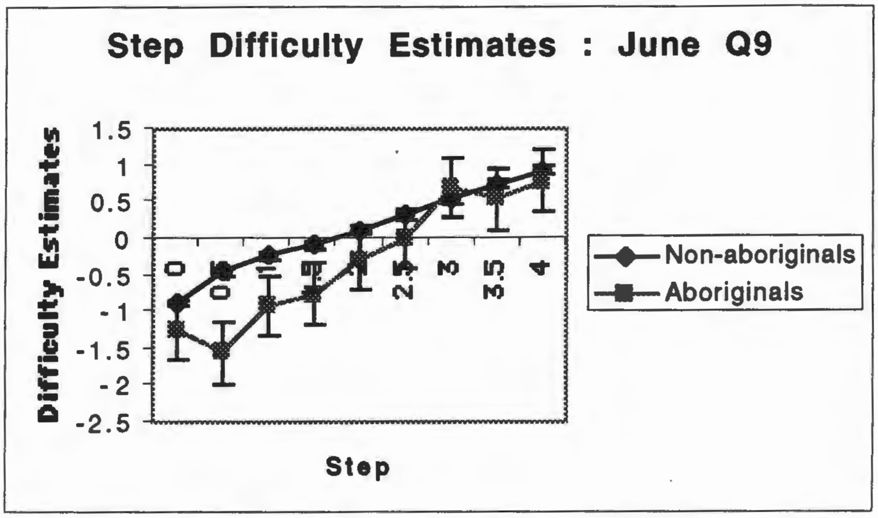

Figure 5. Difficulty estimates for the increments. Note that the aboriginals found the item easier at the lower increments even though overall the non-aboriginals found the item as a whole easier. The difficulty estimates do not overlap for the lower increments but do for the higher increments. For the two plots, the area is large for the lower increments and small for the higher increments.

The partial credit analysis provided a better match for the response sets of the students. Items from the open-ended component that were earlier identified as displaying differential item function were found not to display differential item functioning using this alternate measurement model. The open-ended components for the two administrations of the examination are virtually DIF free.

Chapter Four

Discussion

Obtaining the data set for the responses for both the January and June sitting of the provincial examination provided the opportunity to examine the results for the students, the items and the replicability of the items. The performance of the groups of students was only indirectly the subject of study. Their performance on the items was the subject of our analyses. Three variables, gender, district size and minority group membership, were used to generate comparisons for our specific comparison groupings. The functioning of specific items influences the quality of the evaluative process. Any evaluative process is flawed if the items unfairly favor one group over another.

Mean scores were determined for the multiple-choice subtest for both sittings. The mean scores reflected the performance of a specific group of individuals that provided a basis for comparison. With respect to gender, males and females had scores that were similar. The differences in scores were not statistically significant. The other two factors showed significant differences in their performance or achievement. The larger the school district, the better the performance. However, after considering the effect size, the difference was not important. In contrast, aboriginals scored statistically significantly lower than what non-aboriginals and the effect size was important.

Mean scores were also determined for the open-ended subtest for both sittings. With respect to gender, nonsignificant differences were found. Males and females scored similarly on the open-ended component. The other two variables did not behave like the gender variable. Larger school districts did better than did the smaller districts. Considering aboriginality, non-aboriginals outperformed their aboriginal counterparts. Significant differences were found but only the aboriginality variable had a large effect size. These results were repeated over the two subtests.

All students who wrote the examination were self-selecting. Chemistry 12 is an elective course. The comparisons that are made in this study refer only to those select participants who wrote the final examination. As a Chemistry teacher, it is a known practice that students who are not succeeding are encouraged not to participate in the final examination or they choose not to participate in the final examination. Many of these students are withdrawn from the course. By excluding these students, there is a loss of data that describes our subgroups. The interpretations need to be tempered. These findings are exclusive to our participants.

The results showed that there were nonsignificant differences for two of the variables examined. Males and females that take Chemistry 12 are equal in performance as measured by these two examinations. Students from small, medium, or large school districts performed equally well as measured by the same two examinations. Aboriginals and non-aboriginals are not equal in performance. This information now can now be interpreted. Gender and locale of the student do not have a bearing on performance on the examination. Belonging to an aboriginal group does have a bearing on performance. There is a bias that affects the performance of aboriginals. For the most part, educators can be pleased with these results and realize that more work must be done to improve the performance of aboriginals.

Mean score analysis simply shows the performance of a select group on the multiple choice or the open-ended component. Even though they may have curriculum overlap, comparisons are not truly warranted since the items that comprise each subtest differ in their prescribed learning outcomes. The content is valid and follows the table of specifications that is used to create the test. Scores could be used to gauge the difficulty of the subtest in comparison to other similar subtests. The scores for each of the comparison groups for the January and June sittings were not significantly different so the creators of these tests should be pleased that the two test forms are indeed parallel in terms of difficulty.

DIF Analysis

The comparison of only the mean scores cannot be relied upon to give an accurate assessment of the groups if there are substantial numbers of items that show a bias to one group or another. The mean score analysis is flawed because of differences in-group characteristics that are susceptible to these biases. Alternative methods of analysis are required to check on the performance of the items. Then and only then can valid conclusions be made about the performance of any group. The examination of these items for DIF was the focus of this study.

The Rasch model provides the theoretical foundation for our analysis. For the Rasch model, there is only one item parameter to estimate, that is, item difficulty (MacMillan, 2002). This greatly simplifies the conceptualization of differential item functioning. There is only a test of differences of the item difficulty estimates. The mathematical simplicity of the Rasch model means stable estimates can be obtained for the sample sizes available in this study.

The difficulty estimates generated for the items now shift the emphasis from the performance of the group to the performance of the item. According to the Rasch model, items that perform well will show invariance across groups. Those items can be validated if they discriminate on the basis of ability only. Poorly performing items can be identified and subject to further analysis if so chosen. Items were found in this analysis that displayed DIF.

Difficulty estimates were generated for each element with a comparison group. These estimates were used to identify items that displayed DIF. The items that displayed DIF were then tabulated for both groups of the investigation and the net DIF was determined. Further investigation checked into the replication of DIF by content area and the specific learning outcome. The open-ended items were further investigated using the partial credit analyses to see if DIF occurred at differing levels of performance within the item. Those results provided information not only on how the item performed but also how the markers interacted with the responses.

The multiple-choice components of both sittings were found to contain items that showed DIF. In the forty-eight item component, as few as fourteen items and as many as thirty items showed DIF. It should be noted however that the comparisons involving the aboriginals had the largest error estimates due to their smaller numbers and consequently the fewest items displaying DIF. This portion of the investigation was found to be insensitive because of the amount of error associated with the measures.

Because DIF has been detected in the items that comprise the subtest, the net DIF must be considered. Ideally, having DIF-free items make up the subtest is the ultimate goal. A net DIF of zero would also be acceptable. The DIF is negated over the subtest and the advantage gained by one group over another is neutralized. The presence of DIF in a subtest depends on the combination of items that comprised that sitting. The net DIF ranged from zero to three. Considering that the item count is forty-eight, the amount of DIF is small and should be considered insignificant.

Replication studies are difficult to perform when net DIF is observed. Once again, the uniqueness of each subtest prohibits replication analysis. Multiple choice items over the two subtests shared similar enumeration but not the same specific prescribed learning outcome. Items reused could be checked for DIF if the items were deemed as exhibiting DIF and their performance was monitored. From the comparisons of net DIF, it was observed that the large school districts had a slight advantage over the small school districts over both sittings. Considering all the other comparison groupings, the results did not show any replication between the two sittings. The results were either a reversal of the net DIF or going from a net DIF of zero to showing a net DIF.

The open-ended components of both sittings were found to contain items that showed DIF. The January sitting contained nine items. DIF was detected in as few as one item and as many as eight items. The June sitting contained eleven items. DIF was detected in as few as four items and as many as ten items. The identification of such numbers showing DIF provides an indication of the overall difficulty of the open-ended component. These findings suggest that the June sitting open-ended component was found to be more difficult compared to the January sitting open-ended component for our comparison groupings.

DIF was found in the open-ended component of both sittings; therefore net DIF needs to be considered. Depending on the comparison groups, net DIF ranged from zero to three. The open-ended component consists of a small number of items with mark values that range from two to five. Having a net DIF of three should be reason for concern. Further investigation is warranted to investigate the implications of the specific items and their assigned mark value.

Each item of the open-ended component has a specific prescribed learning outcome. It is rare that items from different sittings would have a common prescribed learning outcome. There may be curricular overlap and that could be the subject of future DIF studies. One item showed DIF that replicated with the same prescribed learning outcome. Other items had dissimilar prescribed learning outcomes but checking for replication is not practical.

Certain patterns of net DIF were observed over the two sittings. For both sittings, females had more items favoring them than did the males. In this case, the net DIF replicated. This needs to be further investigated to see if this is a trend. The other comparisons saw the advantage change between the sittings. Other observations included changing from no advantage or no net DIF to an advantage. The reverse also occurred. These later observations reflect the uniqueness of each subtest. Checking for replication would be a difficult investigation.

The partial credit analyses for differential item functioning showed that specific items in the open-ended component displayed DIF. Those items that were DIF free over the increments awarded for the item had similar difficulty estimates for the two comparison groups. These items were objective in nature and an external factor of a marker was added. The similar difficulty estimates indicate that the marking rubric was consistently applied to the subgroups by the various markers that participated. The marking process for the examination is a strength of the evaluative process that is set in place. DIF cannot be attributed to the interpretations of the markers.

The DIF that was observed through partial credit analysis was limited to a few instances. This is contradictory to the difficulty estimates that were used to check for DIF. The earlier open-ended analysis showed more items displaying DIF. The estimates using the partial credit model are more suited to the identification of DIF; the partial credit model includes difficulty estimates for scoring zero as well as all the incremental scores for that particular item. The FACETS' estimates that were used gave only the overall item difficulty. This is a problem of minimal importance for the few cases of either high or low ability. These estimates show DIF but do not fit the response data as well as the estimates arrived using the partial credit model. Difficulty estimates using BIGSTEPS are more thorough and reinforce the practice of students showing complete solutions. Part marks are available and the utilization of difficulty estimates should coincide with that scheme of arriving at marks. The findings here negate the findings of net DIF found for the open-ended component items that was identified earlier.

The effects of net DIF over the two subtests that comprise the examination also needs to be considered. Since the subtests have different mark values, net DIF cannot be assessed unless the net DIF favors the same group over the two components. It is not sufficient to say that the DIF is negated if the multiple choice component favors one group and the open-ended component favors the other group. The tracking of items that display DIF would be necessitated if net DIF continues to occur favoring one group over another. Those combinations of items should not occur again on future subtests. The testing procedure needs to use items whose characteristics are known and reliable and free of DIF.

There is no interplay between means scores and DIF. An examination of the means scores for the subtests showed that there were significant differences with small effect sizes except for the comparison involving minority group membership. With trivial effect sizes, comparisons are not merited. With minimal amounts of DIF, the comparison of means is not complicated. When it comes to gender, females and males are equal in ability for those who took the examination. With respect to the district size, no differences are found. Using aboriginality as the group of minority membership, the differences are significant with little DIF problems. The difference in performance is statistically significant and merits further investigation.

Implications

The goal of DIF analyses is to check the performance of items. These items are not to interact with the characteristics of any one group. The evaluative process needs to be scrutinized as the process was set up to ensure that students are to be treated fairly when applying for admission to universities and other post-secondary institutions. This fair treatment also provides fuel for the argument that the education experiences may be one of the characteristics of the group with which items are interacting. The three variables of study are based on the premise that the educational experiences are similar. If this premise were incorrect, the test performance differences would be due to differences in the educational experiences.

For the students and teachers, the DIF analysis has found the majority of items function well. Regardless of their gender, locale or aboriginal status, students with equal ability have the same probability of answering the questions correctly. The formal evaluation is a valid process that will determine a portion of their overall grade. These students will be treated fairly based on the inferences made from their scores.

For researchers, finding DIF validates research done in the field of measurement. A DIF analysis provides valuable information on the performance of items that is far superior to that of classical test theory. The DIF analysis gives the researcher another technique for quality control. The relative lack of DIF found here validates the testing process. Since such large emphasis has been placed on this formal testing procedure, its integrity needs to be maintained. The DIF analysis done by researchers serves to validate the process or provides the impetus to investigate the causes of DIF and ultimately promote education change.

For the Ministry of Education that controls education within the province, DIF analysis provides a tool to validate the evaluative process. Mean score analysis is compromised by subgroup characteristics. The process needs to be free of bias and perceived to be fair. Use of the item response model, more specifically the Rasch model, generates difficulty estimates based on solely the ability of the participants. By eliminating other factors, differences can be attributed to group membership. The finding of items that display DIF should be alarming. Since this evaluative process is ongoing, item characteristics should be determined and those that do not perform fairly should not be used. The governing body to ensure that evaluative process is fair should advocate DIF analysis. Other large-scale assessments already have DIF analyses in place. A Rasch based DIF procedure can be used as a gauge to see if measures implemented by the governing body are effective to reduce the amount of item bias in high stakes measures.

The quality of the items and their ability to discriminate accurately are the cornerstones of measurement and assessment. Simply put, DIF analysis validates the examination process. Formal summative evaluations need to have a process in which testing practices can be monitored and corrective measures may result from the information gathered. The two sittings of provincial examinations are by no means perfect instruments of assessment. Instead, it shows varying levels of DIF for both its sub-test components. The amount of DIF that is acceptable is subject to debate and that debate is not the intent of this research. Overall, the DIF analyses demonstrated that the provincial examination, even though it possesses items that show DIF, is a valid instrument of measurement.

From this examination of the items of these provincial examinations, the controlling body can infer the quality of the educational opportunities afforded their students. Because of DIF or the lack of significant DIF, the argument can be made that the examination system works or there are inequities. Change can result if there is a need be to minimize the DIF that is seen and it can be monitored over time to check on the effectiveness of the changes made. Funding practices can be examined since the controlling body funds education. Currently, the funding does not take into consideration where the student is enrolled. In light of the funding for education, changes could be made if need be to enhance the educational experience for those who are disadvantaged due to their school organization as a result of their locale. The quality of the learning experience can be monitored.

The existence of DIF is problematic. Being DIF free would be a desired trait for any type of formal summative evaluation. If DIF occurs, the net DIF needs to be zero. If there is DIF, measures need to be taken to minimize the DIF to reverse this finding. The favoring of one subgroup over another is not tolerated and frowned upon in our society. If DIF is present, measures must be taken to gain a positive result from a negative situation. DIF analyses can spur educational reform. DIF certainly can act as an index that educators can use to evaluate the state of education by ensuring that quality measures are obtained.

References

Allen, N. L. & Wainer, H. (1989). *Nonresponse in declared ethnicity and the identification of differentially functioning items. Program Statistics Research, technical Report No. 898-89.* Princeton, NJ: Educational Testing Service.

Assessment Systems Corporation. (1988). *User's manual for the MicroCat testing system. (Version 3).* St. Paul, MN: Author.

Bidwell, C.E., & Kasarda, J.D. (1975). School district organization and student achievement. *American Sociological Review, 40(February),* 55-70.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals, Handbook I: Cognitive domain.* New York: Longman, Green.

Borich, G., & Tombari, M. (1995). *Educational psychology: A contemporary approach.* New York: Harper Collins College Publishers.

Cameron, I. (1990). Student achievement among native students in British Columbia. *Canadian Journal of Native Education, 17*(1).

Campbell, P. B., & Storo, J. N. (1996). *Girls are . . . boys are . . . myths, stereotypes and gender differences. Math and science for the coed classroom.* Newton: MA: Educational Development Center, Inc.

Clauser, B. E., & Mazor, K.M. (1998). *Using statistical methods and content review for identifying differential item functioning.* Educational Measurement: Issues and Practice, 17, 31-44.

Coleman, P., & Laroque, L. (1986). The small school district in British Columbia: The myths, the reality, and some policy implications. *The Alberta Journal of Educational Research, 32(4),* 323-335.

Dillon, J. T. (1982). Male-female similarities in class participation. *Journal of Educational Research, 75,* 350-353.

Draba, R. E. (1977). The identification and interpretation of item bias. *Research Memorandum No. 25.* Chicago: MESA Psychometric Laboratory.

Edge. J., Martin, C., & Morris, M. (1997). *Promoting gender equity within the classroom.* Chicago: Saint Xavier University, M. A. Action Research Project.

Fowler, W. J. Jr. (1992). *What do we know about school size? What should we know?* Paper presented at the annual general meeting of the American Educational Research Association, San Francisco.

Gambell, T., & Hunter, D. (1997, June). *Leveling the Gender Playing Field? Opportunity and Outcome in Canadian Literacy.* Paper presented at XXV (25th) Annual Conference of the Canadian Society for the Study of Education, St. John's, NF.

Good, T., & Brophy, J. (1991). *Looking into classrooms* (5th ed). New York: Harper Collins Publishers.

Haggerty, S. M. (1991). Gender & school science: Achievement & participation in Canada. *The Alberta Journal of Educational Research, 37*, 195-208.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.Academic.

Hambleton, R. K., & Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park: Sage Publications.

Hanson, R. A., & Farrell, D. (1995). Long term effects on high school seniors of learning to read in kindergarten. *Reading Research Quarterly, 30*(4).

Hativa, N. (1989). Socioeconomic status, aptitude, and gender differences in CAI gains of arithmetic. *Journal of Educational Research, 83*, 11-21.

Hoff Sommers, C. (2000). *The war against boys: How misguided feminism is harming our young men.* New York: Simon & Schuster Books.

Howley, C. B. (1994). *The academic effectiveness of small scale schooling (an update).* ERIC Clearinghouse on Rural Education and Small Schools. (ERIC Document Reproduction Service No. ED 372 897)

Howley, C. B., & Harmon, H. L. (Eds.). (2000). *Small high schools that flourish: Rural Context, case studies, and resources.* Charleston, WV: Appalachia Educational Laboratory.

Indian and Northern Affairs Canada (1997). *Socio-economic indicators in Indian reserves and comparable communities 1971-1991.* Ottawa: unknown.

Lee, V. E., & Smith, J. B. (1996). *High school size: Which works best, and for whom?* Paper presented at the annual general meeting of the American Educational Research Association, New York.

Linacre, J. M. (1996). *A users guide to FACETS.* Chicago: MESA Press.

Linacre, J. M., & Wright, B. D. (1995). *A user's guide to BIGSTEPS: Rasch-Model computer program.* Chicago: MESA Press.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*, 159-173.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Ma, X. (1995). Gender differences in mathematics achievement between Canadian and Asian educational systems. *Journal of Educational Research, 89*, 118-127.

MacMillan, P. D. (2002). *Differential item functioning (DIF) of CBM mathematics probes: An application of a many faceted Rasch Poisson model.* Paper presented at Canadian Society for Studies in Education, Toronto, ON.

Malatest, R., Barry, J., Krebs, S., & Whyte, K. (2002). *Parent and education engagement partnership project.* Victoria, BC: R. A. Malatest & Associates Ltd.

Marx, R., & Grieve, T. (1988). *The learners of British Columbia (Commissioned Papers: Volume 2).* Victoria: British Columbia Royal Commission on Education.

McGivern, R. F., Huston, J. P., Byrd, D., King, Tl, Siegle, G. J., & Reilly, J. (1997). Sex differences in visual recognition: Support for a sex related difference in attention in adults & children. *Brain & Cognition, 34,* 323-336.

Monk, D. H. (1992). *Modern conceptions of educational quality and state policy regarding small schooling units. Source book on school and district size, cost, and quality.* Minneapolis, Hubert Humphrey Institute of Public Affairs and North Central Regional Educational Lab: 35-49.

More, A. J. (1998). *Ways of learning, learning styles and First Nations students: A teacher resource.* Vancouver: University of British Columbia.

Pashley, P. J. (1992). *Graphical IRT-based DIF analyses.* Princeton, NJ.: Educational Testing Service.

Pope, G. A. (1998). *Nonparametric item response modelling and gender differential functioning (DIF) analysis of the Eysenck Personality Questionnaire (EPQ).* Unpublished master's thesis, University of Northern British Columbia, Prince George, British Columbia, Canada.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14(2),* 197-207.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Reyhner, J. (1992). American Indians out of school: A review of school-based causes and solutions. *Journal of American Indian Education.* May 1992.

Riles, S. B. (1995). *High school completion rates by Native Americans.* Portland: Northwest Regional Educational Lab.

Roellke, C. (1996). *Curriculum adequacy and quality in high schools enrolling fewer than 400 pupils (9-12).* ERIC Digest: ERIC Clearinghouse on Rural Education and Small Schools, Charleston, VW. Office of Educational Research and Improvement.

Scheuneman, J. D. & Slaughter, C. (1991). *Issues of test bias, item bias, and group Differences and what to do while waiting for answers.*

Schofield, H. L. (1982). Sex, grade level, and the relationship between mathematics attitude and achievement in children. *Journal of Educational Research, 75,* 280-284.

The Women's Freedom Network. (1998, July). *The myth that schools shortchange girls: Social science in the service of deception.* Washington, DC: Judith Kleinfeld.

Tocci, C., & Engelhard, G. (1991). Achievement, parental support & gender differences in attitudes toward mathematics. *Journal of Educational Research, 84,* 280-286.

Walberg, H. J. (1992). *On local control: Is bigger better? Source book on school district Size, cost, and quality.* Minneapolis, MN: Hubert Humphrey Institute of Public Affairs.

Walberg, H. J., & Fowler, W. J. (1987). Expenditure and size efficiencies of public school districts. *Educational Researcher, 16*: 5-15.

Weiss, D. J., & Yoes, M. E. (1991). *Item response theory.* Boston: Kluwer Academic

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Wright, B. D., Mead, R., & Draba, R. (1976). Detecting and correcting test item bias with a logistic response model. *MESA Research Memorandum Number 22.* [On-line] Available: http://www.rasch.org/memo22.html

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendix A

School District Size Characterization

British Columbia's fifty-nine school districts are characterized by the size of the centre where the school board offices are located. Those centres whose population is greater than 100 000 are classified as large school districts while those centres whose population is less than 25 000 are classified as small districts. Those centres whose population is intermediate are classified as medium districts.

Table 1

School District Size Characterization

School District Number	Region	Size
5	Southeast Kootenay	Small
6	Rocky Mountain	Small
8	Kootenay Lake	Small
10	Arrow Lakes	Small
19	Revelstoke	Small
20	Kootenay-Columbia	Small
22	Vernon	Small
23	Central Okanagan	Medium
27	Cariboo-Chilcotin	Small
28	Quesnel	Small
33	Chilliwack	Medium
34	Abbotsford	Medium
35	Langley	Medium
36	Surrey	Large
37	Delta	Medium
38	Richmond	Large
39	Vancouver	Large
40	New Westminster	Large

41	Burnaby	Large
42	Maple Ridge/Pitt Meadows	Medium
43	Coquitlam	Large
44	North Vancouver	Large
45	West Vancouver	Large
46	Sunshine Coast	Small
47	Powell River	Small
48	Howe Sound	Small
49	Central Coast	Small
50	Haida Gwaii/Queen Charlotte	Small
51	Boundary	Small
52	Prince Rupert	Small
53	Okanagan Similkameen	Small
54	Bulkley Valley	Small
57	Prince George	Medium
58	Nicola-Similkameen	Small
59	Peace River South	Small
60	Peace River North	Small
61	Greater Victoria	Large
62	Sooke	Small
63	Saanich	Small
64	Gulf Islands	Small
67	Okanagan Skaha	Small
68	Nanaimo/Ladysmith	Medium
69	Qualicum	Small
70	Alberni	Small
71	Comox Valley	Small

72	Campbell River	Small
73	Kamloops/Thompson	Medium
74	Gold Trail	Small
75	Mission	Small
78	Fraser-Cascade	Small
79	Cowichan Valley	Small
81	Fort Nelson	Small
82	Coast Mountains	Small
83	North Okanagan-Shuswap	Small
84	Vancouver Island West	Small
85	Vancouver Island North	Small
87	Stikine	Small
91	Nechako Lakes	Small
92	Nisga'a	Small
101	Yukon	Small

Appendix B

Examination Specifications

Each examination examines curriculum as established by the provincial body. The examinations are similar in structure and value. The emphasis for each of the content areas is outlined by the Prescribed Learning Outcome (PLO).

Table B1
Multiple Choice Composition for January 1999

Item	Cognitive Level	Content Organizer	PLO	Item	Cognitive Level	Content Organizer	PLO
1	Understanding	Kinetics	A3	25	Knowledge	Acid/Base	L11
2	Knowledge	Kinetics	B3	26	Knowledge	Acid/Base	L10
3	Knowledge	Kinetics	B2	27	Understand	Acid/Base	L6
4	Higher Mental	Kinetics	B4	28	Understand	Acid/Base	L11
5	Understanding	Kinetics	B6	29	Higher Mental	Acid/Base	K9/L11
6	Understanding	Kinetics	C5	30	Understanding	Acid/Base	N3
7	Understanding	Equilibria	D4	31	Understanding	Acid/Base	P1
8	Understanding	Equilibria	E2	32	Knowledge	Acid/Base	Q3
9	Understanding	Equilibria	E2/E5	33	Knowledge	Acid/Base	R1
10	Higher Mental	Equilibria	E3	34	Understanding	Acid/Base	O5
11	Knowledge	Equilibria	F3	35	Understanding	Acid/Base	K11
12	Higher Mental	Equilibria	F4	36	Understanding	Acid/Base	P5
13	Understanding	Solubility	F7	37	Understanding	Redox	S1
14	Knowledge	Solubility	G4	38	Understanding	Redox	S2
15	Understanding	Solubility	H2	39	Understanding	Redox	S5
16	Understanding	Solubility	H3	40	Understanding	Redox	S6
17	Knowledge	Solubility	I2	41	Understanding	Redox	T3
18	Understanding	Solubility	I3	42	Higher Mental	Redox	T4
19	Understanding	Solubility	I5	43	Understanding	Redox	U3/U4
20	Understanding	Solubility	H1/I4	44	Knowledge	Redox	U8

21	Knowledge	Acid//Base	J7	45	Knowledge	Redox	V2
22	Higher Mental	Acid/Base	H5	46	Knowledge	Redox	V3
23	Higher Mental	Acid/Base	K1	47	Understanding	Redox	W4
24	Understanding	Acid/Base	K6	48	Knowledge	Redox	W5

Table B2

Written Response Composition for January 1999

Item	Cognitive Level	Item Value	Content Organizer	PLO
1	Understanding	3	Kinetics	C2
2	Understanding	5	Equilibria	D4, F1, F5
3	Understanding	4	Solubility	G5, I3
4	Understanding	4	Acid/Base	M4, N1, N3
5	Understanding	4	Acid/Base	M3
6	Understanding	4	Acid/Base	P1, P4, P6
7	Understanding	3	Redox	T2
8	Understanding	3	Redox	T6
9	Understanding	2	Redox	W6

Table B3

Multiple Choice Composition for June 1999

Item	Cognitive Level	Content Organizer	PLO	Item	Cognitive Level	Content Organizer	PLO
1	Knowledge	Kinetics	A2	25	Knowledge	Acid/Base	L1
2	Higher Mental	Kinetics	A2	26	Knowledge	Acid/Base	L3
3	Understanding	Kinetics	A6	27	Higher Mental	Acid/Base	L4
4	Understanding	Kinetics	B6	28	Understanding	Acid/Base	K5/J8
5	Higher Mental	Kinetics	B3/B9	29	Understanding	Acid/Base	L12

6	Knowledge	Kinetics	C3	30	Understanding	Acid/Base	M1/N4
7	Understanding	Equilibria	D7	31	Understanding	Acid/Base	N3
8	Understanding	Equilibria	E2	32	Knowledge	Acid/Base	O5
9	Understanding	Equilibria	E2/E5	33	Knowledge	Acid/Base	O2
10	Understanding	Equilibria	E4	34	Understanding	Acid/Base	P5
11	Understanding	Equilibria	F2	35	Knowledge	Acid/Base	Q1
12	Higher Mental	Equilibria	F4	36	Knowledge	Acid/Base	R1
13	Understanding	Solubility	F7	37	Understanding	Redox	S1
14	Understanding	Solubility	G8	38	Understanding	Redox	S2
15	Understanding	Solubility	H1	39	Understanding	Redox	S2
16	Understanding	Solubility	H7	40	Understanding	Redox	S6
17	Understanding	Solubility	I3	41	Knowledge	Redox	T1
18	Knowledge	Solubility	I6	42	Understanding	Redox	T4
19	Higher Mental	Acid/Base	J1	43	Knowledge	Redox	V2
20	Understanding	Acid/Base	J7	44	Understanding	Redox	U10
21	Understanding	Acid/Base	J8	45	Understanding	Redox	U2
22	Understanding	Acid/Base	K1	46	Knowledge	Redox	U11
23	Understanding	Acid/Base	K6	47	Understanding	Redox	W4
24	Knowledge	Acid/Base	K11	48	Knowledge	Redox	W1

Table B4

Written Response Composition for June 1999

Item	Cognitive Level	Item Value	Content Organizer	PLO
1	Understanding	3	Kinetics	B9
2	Understanding	4	Equilibria	D3, D4, F5
3	Knowledge	2	Equilibria	E2

4	Understanding	2	Solubility	H3
5	Understanding	4	Solubility	I4
6	Understanding	2	Acid/Base	K7
7	Understanding	4	Acid/Base	M3, M4, M5
8	Understanding	3	Acid/Base	P3
9	Understanding	4	Redox	T6
10	Understanding	2	Redox	U1, U7
11	Higher Mental	2	Redox	W4

Appendix C

Difficulty Estimates

Each item from the multiple-choice component was analyzed. Difficulty estimates and the associated error of estimate are tabulated below.

Insert Table C1

Insert Table C2

Table C1

Difficulty Estimates for January 1999 Multiple Choice Component

Item	Gender				District Size						Aboriginality			
	Female		Male		Small		Medium		Large		Non-aboriginal		Aboriginal	
	D	SE	D	SE	D	SE	D	SE	D	SE	D	SE	D	SE
1	0.22	(0.05)	-0.06	(0.06)	0.29	(0.07)	0.04	(0.08)	-0.04	(0.06)	0.08	(0.04)	-0.26	(0.53)
2	-1.80	(0.09)	-2.30	(0.10)	-2.09	(0.12)	-2.55	(0.16)	-0.79	(0.09)	-2.02	(0.07)	deleted	deleted
3	-0.51	(0.06)	-0.34	(0.06)	-0.65	(0.08)	-0.36	(0.09)	-0.31	(0.06)	-0.43	(0.04)	0.01	(0.52)
4	0.98	(0.05)	1.03	(0.05)	1.00	(0.07)	1.08	(0.08)	0.97	(0.06)	1.00	(0.04)	1.05	(0.53)
5	-1.57	(0.08)	-1.66	(0.08)	-1.48	(0.10)	-1.76	(0.12)	-1.64	(0.09)	-1.62	(0.06)	-1.72	(0.76)
6	-1.18	(0.07)	-1.13	(0.07)	-1.17	(0.09)	-1.20	(0.10)	-1.12	(0.07)	-1.15	(0.05)	-2.48	(1.02)
7	-0.22	(0.06)	-0.62	(0.06)	-0.16	(0.07)	-0.34	(0.09)	-0.63	(0.07)	-0.41	(0.04)	-1.24	(0.65)
8	-0.56	(0.06)	-0.60	(0.06)	-0.50	(0.08)	-0.49	(0.09)	-0.68	(0.07)	-0.58	(0.04)	-0.86	(0.59)
9	0.04	(0.05)	-0.12	(0.06)	-0.05	(0.07)	0.06	(0.08)	-0.08	(0.06)	-0.04	(0.04)	-0.26	(0.53)
10	2.44	(0.07)	2.52	(0.07)	2.43	(0.08)	2.52	(0.10)	2.50	(0.07)	2.49	(0.05)	1.34	(0.56)
11	-0.33	(0.06)	-0.33	(0.06)	-0.43	(0.08)	-0.39	(0.09)	-0.23	(0.06)	-0.33	(0.04)	-0.54	(0.55)
12	0.30	(0.05)	0.11	(0.06)	0.22	(0.07)	0.25	(0.08)	0.18	(0.06)	0.21	(0.04)	0.52	(0.51)

13	-0.89 (0.07)	-0.55 (0.06)	-0.62 (0.08)	-0.57 (0.09)	-0.86 (0.07)	-0.72 (0.04)	-0.54 (0.55)
14	-0.53 (0.06)	-0.46 (0.06)	-0.40 (0.07)	-0.62 (0.09)	-0.50 (0.06)	-0.49 (0.04)	-0.86 (0.59)
15	-1.62 (0.08)	-1.44 (0.08)	-1.56 (0.10)	-1.70 (0.12)	-1.42 (0.08)	-1.54 (0.06)	-0.86 (0.59)
16	0.11 (0.06)	0.21 (0.06)	0.27 (0.07)	-0.08 (0.08)	0.20 (0.06)	0.15 (0.04)	1.05 (0.53)
17	-1.24 (0.07)	-0.70 (0.06)	-1.15 (0.09)	-1.02 (0.10)	-0.79 (0.07)	-0.95 (0.05)	-0.54 (0.55)
18	0.58 (0.05)	0.76 (0.05)	0.68 (0.07)	0.86 (0.08)	0.57 (0.06)	0.67 (0.04)	0.78 (0.52)
19	-0.96 (0.07)	-0.81 (0.06)	-0.85 (0.08)	-0.82 (0.09)	-0.95 (0.07)	-0.89 (0.05)	0.01 (0.52)
20	1.03 (0.05)	0.96 (0.05)	0.98 (0.07)	0.90 (0.08)	1.05 (0.06)	1.00 (0.04)	0.26 (0.51)
21	-1.38 (0.07)	-1.36 (0.07)	-1.46 (0.10)	-1.28 (0.11)	-1.36 (0.08)	-1.37 (0.05)	-1.72 (0.76)
22	2.66 (0.07)	2.51 (0.06)	2.61 (0.09)	2.73 (0.11)	2.50 (0.07)	2.58 (0.05)	2.55 (0.76)
23	1.77 (0.06)	1.84 (0.06)	1.75 (0.07)	1.50 (0.08)	2.00 (0.06)	1.82 (0.04)	0.26 (0.51)
24	-0.43 (0.06)	-0.36 (0.06)	-0.37 (0.07)	-0.46 (0.09)	-0.38 (0.06)	-0.40 (0.04)	-0.86 (0.59)
25	-0.09 (0.06)	-0.03 (0.06)	0.00 (0.07)	0.00 (0.08)	-0.13 (0.06)	-0.07 (0.04)	0.78 (0.52)
26	0.75 (0.05)	0.86 (0.05)	1.00 (0.07)	1.01 (0.08)	0.58 (0.06)	0.80 (0.04)	1.05 (0.53)
27	1.75 (0.06)	1.60 (0.06)	1.84 (0.07)	1.67 (0.09)	1.58 (0.06)	1.68 (0.04)	0.78 (0.52)

28	0.06	(0.06)	0.27	(0.06)	0.28	(0.07)	0.12	(0.08)	0.10	(0.06)	0.16	(0.04)	0.78	(0.52)
29	0.66	(0.05)	0.42	(0.05)	0.59	(0.07)	0.40	(0.08)	0.58	(0.06)	0.54	(0.04)	0.52	(0.51)
30	1.00	(0.05)	1.22	(0.05)	1.06	(0.07)	1.21	(0.08)	1.10	(0.06)	1.11	(0.04)	1.05	(0.53)
31	-0.91	(0.07)	-0.88	(0.07)	-0.92	(0.08)	-0.85	(0.10)	-0.90	(0.07)	-0.89	(0.05)	-1.72	(0.76)
32	-0.42	(0.06)	-0.30	(0.06)	-0.56	(0.08)	-0.36	(0.09)	-0.23	(0.06)	-0.36	(0.04)	0.01	(0.52)
33	-0.85	(0.06)	-0.91	(0.07)	-1.17	(0.09)	-0.80	(0.09)	-0.74	(0.07)	-0.88	(0.05)	-0.54	(0.55)
34	-1.65	(0.08)	-1.86	(0.09)	-1.79	(0.11)	-1.64	(0.12)	-1.79	(0.09)	-1.75	(0.06)	deleted	deleted
35	-0.71	(0.06)	-0.74	(0.06)	-0.70	(0.08)	-0.77	(0.09)	-0.72	(0.07)	-0.72	(0.04)	-1.24	(0.65)
36	0.83	(0.05)	0.43	(0.05)	0.62	(0.07)	0.96	(0.08)	0.48	(0.06)	0.63	(0.04)	0.78	(0.52)
37	0.19	(0.05)	0.29	(0.06)	0.31	(0.07)	0.29	(0.08)	0.16	(0.06)	0.23	(0.04)	0.78	(0.52)
38	0.37	(0.05)	0.33	(0.05)	0.49	(0.07)	0.20	(0.08)	0.33	(0.06)	0.35	(0.04)	-0.26	(0.53)
39	-1.55	(0.08)	-1.48	(0.08)	-1.54	(0.10)	-1.58	(0.12)	-1.46	(0.08)	-1.52	(0.06)	-0.86	(0.59)
40	-0.93	(0.07)	-0.75	(0.06)	-0.92	(0.08)	-1.00	(0.10)	-0.71	(0.07)	-0.85	(0.05)	-0.26	(0.53)
41	0.12	(0.06)	0.36	(0.05)	0.41	(0.07)	0.17	(0.08)	0.15	(0.06)	0.23	(0.04)	1.05	(0.53)
42	0.41	(0.05)	0.46	(0.05)	0.46	(0.07)	0.41	(0.08)	0.43	(0.06)	0.43	(0.04)	0.52	(0.51)

43	0.30	(0.05)	0.36	(0.05)	0.28	(0.07)	0.28	(0.08)	0.37	(0.06)	0.32	(0.04)	0.52	(0.51)
44	1.71	(0.06)	1.46	(0.06)	1.58	(0.07)	1.58	(0.08)	1.58	(0.06)	1.58	(0.04)	1.05	(0.53)
45	1.27	(0.05)	1.22	(0.05)	1.31	(0.07)	1.52	(0.08)	1.08	(0.06)	1.24	(0.04)	1.66	(0.59)
46	-0.73	(0.06)	-0.80	(0.06)	-1.07	(0.09)	-0.80	(0.09)	-0.55	(0.06)	-0.76	(0.05)	-2.48	(1.02)
47	1.36	(0.06)	1.44	(0.06)	1.32	(0.07)	1.66	(0.09)	1.34	(0.06)	1.40	(0.04)	0.52	(0.51)
48	0.13	(0.06)	-0.08	(0.06)	-0.19	(0.07)	0.03	(0.08)	0.17	(0.06)	0.03	(0.04)	0.52	(0.51)
n	1744		1760		1071		773		1059		3492		17	
df	19		19		19		19		19		19		3	

Table C2

Difficulty Estimates for the June 1999 Multiple Choice Component

Item	Gender				District Size						Aboriginality			
	Female		Male		Small		Medium		Large		Non-aboriginal		Aboriginal	
	D	SE	D	SE	D	SE	D	SE	D	SE	D	SE	D	SE
1	-1.15	(0.04)	-1.21	(0.05)	-1.03	(0.06)	-1.27	(0.07)	-1.21	(0.04)	-1.19	(0.03)	-0.36	(0.37)
2	-0.31	(0.04)	-0.30	(0.04)	-0.35	(0.06)	-0.56	(0.06)	-0.21	(0.03)	-0.30	(0.03)	-0.95	(0.40)
3	1.56	(0.04)	1.20	(0.03)	1.17	(0.05)	1.15	(0.06)	1.52	(0.03)	1.38	(0.02)	0.82	(0.38)
4	-0.82	(0.04)	-1.32	(0.05)	-0.95	(0.06)	-0.93	(0.07)	-1.11	(0.04)	-1.04	(0.03)	-1.11	(0.42)
5	2.16	(0.04)	1.85	(0.04)	1.86	(0.06)	2.19	(0.06)	2.00	(0.03)	2.00	(0.03)	1.99	(0.47)
6	-1.68	(0.06)	-1.77	(0.05)	-1.79	(0.08)	-1.80	(0.09)	-1.67	(0.06)	-1.72	(0.04)	-1.30	(0.44)
7	1.39	(0.03)	1.29	(0.03)	1.39	(0.06)	1.60	(0.06)	1.25	(0.03)	1.34	(0.02)	1.42	(0.41)
8	-0.93	(0.04)	-0.95	(0.04)	-0.74	(0.06)	-0.87	(0.07)	-1.05	(0.04)	-0.94	(0.03)	0.50	(0.38)
9	-0.19	(0.04)	-0.26	(0.04)	-0.16	(0.05)	-0.15	(0.06)	-0.27	(0.03)	-0.23	(0.03)	0.03	(0.36)
10	0.01	(0.03)	0.02	(0.04)	-0.02	(0.05)	-0.02	(0.06)	0.03	(0.03)	0.01	(0.03)	-0.23	(0.37)
11	-0.73	(0.04)	-0.44	(0.04)	-0.58	(0.06)	-0.48	(0.06)	-0.63	(0.04)	-0.59	(0.03)	-0.64	(0.38)

12	0.42 (0.03)	0.39 (0.03)	0.57 (0.05)	0.57 (0.05)	0.30 (0.03)	0.41 (0.02)	0.42 (0.36)
13	-1.83 (0.05)	-1.68 (0.05)	-1.52 (0.07)	-1.61 (0.08)	-1.91 (0.05)	-1.76 (0.04)	-0.95 (0.40)
14	0.31 (0.03)	0.36 (0.03)	0.56 (0.05)	0.33 (0.06)	0.26 (0.03)	0.33 (0.02)	0.55 (0.37)
15	-0.90 (0.04)	-0.78 (0.04)	-0.82 (0.06)	-0.94 (0.07)	-0.81 (0.04)	-0.84 (0.03)	-0.79 (0.39)
16	-0.86 (0.04)	-0.69 (0.04)	-0.86 (0.06)	-0.92 (0.07)	-0.70 (0.04)	-0.78 (0.03)	-0.79 (0.39)
17	0.22 (0.03)	0.31 (0.03)	0.34 (0.05)	0.50 (0.05)	0.16 (0.03)	0.26 (0.02)	0.42 (0.36)
18	0.67 (0.03)	0.64 (0.03)	0.79 (0.05)	0.78 (0.05)	0.58 (0.03)	0.66 (0.02)	0.96 (0.38)
19	1.21 (0.03)	1.38 (0.03)	1.06 (0.05)	1.30 (0.06)	1.38 (0.03)	1.30 (0.02)	0.42 (0.36)
20	-1.13 (0.04)	-1.05 (0.04)	-0.96 (0.06)	-1.15 (0.07)	-1.12 (0.04)	-1.09 (0.03)	-1.11 (0.42)
21	-0.75 (0.04)	-0.70 (0.04)	-0.65 (0.06)	-0.77 (0.06)	-0.74 (0.04)	-0.73 (0.03)	-0.10 (0.36)
22	2.40 (0.04)	2.27 (0.04)	2.20 (0.06)	2.32 (0.07)	2.38 (0.04)	2.33 (0.03)	2.23 (0.51)
23	-0.23 (0.04)	-0.10 (0.04)	-0.27 (0.06)	-0.29 (0.06)	-0.09 (0.03)	-0.17 (0.03)	0.03 (0.36)
24	-1.01 (0.04)	-0.77 (0.04)	-0.85 (0.06)	-1.00 (0.07)	-0.87 (0.04)	-0.89 (0.03)	-0.36 (0.37)
25	-0.33 (0.04)	-0.35 (0.04)	-0.46 (0.06)	-0.29 (0.06)	-0.31 (0.03)	-0.34 (0.03)	-0.23 (0.37)
26	-1.18 (0.04)	-1.24 (0.05)	-1.35 (0.07)	-1.30 (0.07)	-1.13 (0.04)	-1.21 (0.03)	-1.50 (0.46)

27	1.08 (0.03)	0.83 (0.03)	1.12 (0.05)	1.06 (0.06)	0.88 (0.03)	0.96 (0.02)	0.82 (0.38)
28	2.46 (0.04)	2.29 (0.04)	2.36 (0.07)	2.36 (0.07)	2.38 (0.04)	2.37 (0.03)	2.23 (0.51)
29	-1.06 (0.04)	-1.07 (0.04)	-0.92 (0.06)	-1.08 (0.07)	-1.12 (0.04)	-1.07 (0.03)	-1.30 (0.44)
30	-0.45 (0.04)	-0.45 (0.04)	-0.52 (0.06)	-0.61 (0.06)	-0.37 (0.03)	-0.44 (0.03)	-1.30 (0.44)
31	0.43 (0.03)	0.48 (0.03)	0.37 (0.05)	0.30 (0.06)	0.53 (0.03)	0.46 (0.02)	-0.23 (0.37)
32	-1.22 (0.04)	-1.60 (0.05)	-1.31 (0.07)	-1.31 (0.07)	-1.45 (0.04)	-1.39 (0.03)	-1.72 (0.49)
33	-0.14 (0.04)	0.12 (0.04)	-0.16 (0.05)	-0.23 (0.06)	0.10 (0.03)	-0.01 (0.03)	-0.36 (0.37)
34	0.77 (0.03)	0.59 (0.03)	0.83 (0.05)	0.84 (0.05)	0.59 (0.03)	0.68 (0.02)	0.42 (0.36)
35	0.74 (0.03)	0.71 (0.03)	0.41 (0.05)	0.55 (0.05)	0.89 (0.03)	0.73 (0.02)	0.42 (0.36)
36	-0.57 (0.04)	-0.51 (0.04)	-0.63 (0.06)	-0.60 (0.06)	-0.49 (0.03)	-0.54 (0.03)	-1.11 (0.42)
37	0.04 (0.03)	0.39 (0.03)	0.24 (0.05)	0.17 (0.06)	0.22 (0.03)	0.21 (0.02)	0.96 (0.38)
38	0.65 (0.03)	0.69 (0.03)	0.73 (0.05)	0.74 (0.05)	0.63 (0.03)	0.67 (0.02)	0.42 (0.36)
39	-0.64 (0.04)	-0.40 (0.04)	-0.47 (0.06)	-0.49 (0.06)	-0.54 (0.04)	-0.52 (0.03)	-0.36 (0.37)
40	-0.79 (0.04)	-0.60 (0.04)	-0.78 (0.06)	-0.73 (0.06)	-0.65 (0.04)	-0.70 (0.03)	-0.23 (0.37)
41	-1.25 (0.04)	-1.39 (0.05)	-1.22 (0.07)	-1.30 (0.07)	-1.36 (0.04)	-1.32 (0.03)	-0.79 (0.39)

42	0.61 (0.03)	0.78 (0.03)	0.69 (0.05)	0.69 (0.05)	0.69 (0.03)	0.69 (0.02)	0.82 (0.38)
43	2.28 (0.04)	2.01 (0.04)	1.83 (0.06)	2.30 (0.07)	2.20 (0.03)	2.14 (0.03)	1.78 (0.45)
44	0.21 (0.03)	0.07 (0.04)	0.03 (0.05)	0.17 (0.06)	0.18 (0.03)	0.14 (0.02)	0.03 (0.36)
45	-0.62 (0.04)	-0.37 (0.04)	-0.36 (0.06)	-0.41 (0.06)	-0.58 (0.04)	-0.50 (0.03)	-0.10 (0.36)
46	0.43 (0.03)	0.29 (0.03)	0.30 (0.05)	0.30 (0.06)	0.40 (0.03)	0.36 (0.02)	0.29 (0.36)
47	0.88 (0.03)	0.93 (0.03)	0.94 (0.05)	0.98 (0.06)	0.87 (0.03)	0.90 (0.02)	1.10 (0.39)
48	-0.16 (0.04)	0.09 (0.04)	-0.08 (0.05)	-0.08 (0.06)	-0.01 (0.03)	-0.04 (0.03)	-0.10 (0.36)
n	4490	4429	1765	1660	5494	8923	35
df	19	19	19	19	19	19	6